EM-алгоритм: оценка фильтров

В этой лекции мы попробуем разобраться, как EM-алгоритм помогает оценивать параметры для модели фильтрации и фильтра Калмана.

5.1 Оценка параметров: вывод правдоподобия

Рассмотрим в деталях модель системы; при этом мы без потери основного смысла существенно упростим вычисления: в оригинальной статье Калмана ещё добавлялись управляющие воздействия, которые вносили смещение в шум. Реальные состояния эволюционируют в соответствии с

$$Y_{t+1} = AY_t + U_{t+1},$$

где Y_t, U_t принимают значения в $\mathbb{R}^n$, при этом U_t – это последовательность шумов. Модель наблюдений тоже линейна и со своим шумом:

$$X_{t+1} = BY_{t+1} + W_{t+1}, \ -$$

про неё мы также предполагаем, что X_t принимает значения в $\mathbb{R}^n$. Матрицы A, B без ограничений, но они должны быть совместимы по размерностям. Про шумы предполагается, что они в совокупности независимы, $U_t \sim \mathcal{N}(0, R_y)$ и $W_t \sim \mathcal{N}(0, R_x)$.

Таким образом, мы имеем вероятностную модель, в которой параметрами θ являются матрицы A, B, ковариационные матрицы векторов шума R_x, R_y, а также параметры стартового состояния $Y_1 \sim \mathcal{N}(\xi, \Lambda)$. Данные включают в себя только набор $X_1, .., X_n$; переменные $Y_1, .., Y_n$ не наблюдаются.

Время EM-алгоритма! Мы, как и раньше, начинаем с того, что записываем функцию

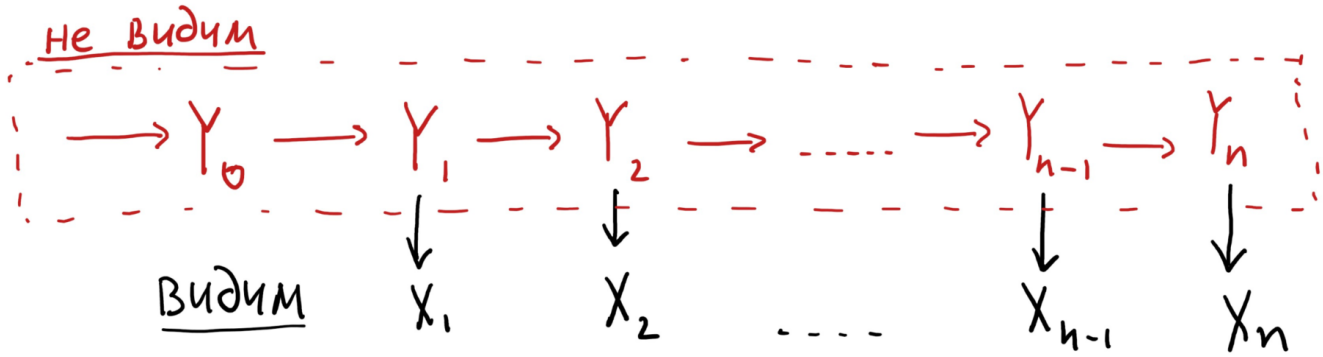
$$Q(\theta^0|\theta) := \mathbb{E}_{\theta^0} \left[\ln p_\theta(X, Y) \mid X\right]$$

и понимаем, что так легко разбить зависимости и выписать условные вероятности, как в случае смеси не получится. Вместо этого мы пойдём более неспешным путём и не будем пока избавляться от матожиданий $\mathbb{E}_{\theta^0} [\cdot \mid X]$.

Начнём с того, что перепишем в точности выражение под знаком матожидания:

$$\ln p_\theta(X, Y) = \ln p_\theta(X|Y) + \ln p_\theta(Y).$$

В данном случае помогает знание зависимостей в величинах X, Y, которые обозначены на рисунке.



Первое слагаемое получается преобразовать как

$$\ln p_\theta(X|Y) = \ln p_\theta(X_1, .., X_{n-1}|X_n, Y) + \ln p_\theta(X_n|Y) = \ln p_\theta(X_1, .., X_{n-1}|Y) + \ln p_\theta(X_n|Y_n)$$

и это приводит к привычной сумме

$$\ln p_\theta(X|Y) = \sum_{i=1}^{n} \ln p_\theta(X_i|Y_i).$$

Что касается второго, скрытый сигнал Y существует по своей динамике и мы видим что-то похожее на случай цепей Маркова с первой недели (и это не случайность):

$$\ln p_\theta(Y) = \ln p_\theta(Y_1) + \sum_{i=1}^{n-1} p_\theta(Y_{i+1}|Y_i).$$

Таким образом, получается лог-правдоподобие из трёх компонент, в каждую из которых входят свои параметры, общих параметров они не имеют:

$$\ln p_\theta(X, Y) = \sum_{i=1}^{n} \ln p_\theta(X_i|Y_i) + \sum_{i=1}^{n-1} p_\theta(Y_{i+1}|Y_i) + \ln p_\theta(Y_1).$$

Первая очень похожа на правдоподобие гауссовской линейной модели регрессии, вторая – на уже ранее виденную модель авторегрессии (но векторной), а третья связана с неопределённостью стартового состояния.

Если мы воспользуемся теперь явным видом распределений и соберём константы в одну общую константу, получим

$$\begin{aligned}
&Q(\theta^0|\theta) = const - \\
&- \frac{n}{2} \ln \det R_x - \frac{1}{2} \mathbb{E}_{\theta_0} \left[\sum_{i=1}^{n} \left((X_i - BY_i)^T R_x^{-1} (X_i - BY_i) \right) \mid X \right] - \\
&- \frac{n}{2} \ln \det R_y - \frac{1}{2} \mathbb{E}_{\theta_0} \left[\sum_{i=1}^{n-1} (Y_{i+1} - AY_i)^T R_y^{-1} (Y_{i+1} - AY_i) \mid X \right] - \\
&- \frac{1}{2} \ln \det(\Lambda) - \frac{1}{2} (Y_1 - \xi)^T \Lambda^{-1} (Y_1 - \xi).
\end{aligned}$$

Здесь очень много всего, но тем не менее есть структура, а если предположить, что все нужные матожидания можно вычислить (мы чуть позже увидим, как именно), то можно явно промаксимизировать по параметрам $\theta = \{A, B, R_x, R_y, \xi, \Lambda\}$. Заметьте, что матожидания берутся по модели данных с параметрами θ^0, поэтому мы можем вносить производные по θ под знак матожидания.

5.2 Оценка параметров: M-шаг

Все вычисления достаточно технические, их можно без проблем проделать самому, если уметь вычислять матричные производные. Нужные формулы вспомним на ходу. Поскольку шумы гауссовские, то оказывается, что можно в правильном порядке провести дифференцирование, приравнять производные к нулю и получить аналитическое точное решение. Хорошей идеей оказывается адресовать параметры авторегрессионной и регрессионной частей отдельно.

Оптимизируем по A. Матрица A входит только в авторегрессионную компоненту; помним, что дифференцирование можно внести под матожидание, потому что матожидание берётся по модели с A^0, а не с A. Смотрим на выражение и понимаем, что здесь нам понадобится знакомая некоторым формула (1)

$$\partial_X \mathrm{tr}\,(AX) = A$$

и более сложная фомула (2)

$$\partial_X \mathrm{tr}\left(AXBX^TC\right) = BX^TCA + B^TX^TA^TC^T.$$

Применим, отбросив не зависящие от A члены:

$$\partial_A Q = \partial_A \mathbb{E}_{\theta^0}\left[\frac{1}{2}\sum_{i=0}^{n-1}\left(Y_{i+1}^T R_y^{-1} Y_{i+1} - Y_i^T A^T R_y^{-1} Y_{i+1} - Y_{i+1} R_y^{-1} A Y_i + Y_i^T A^T R_y^{-1} A Y_i\right) \mid X\right] =$$
$$= \sum_{i=0}^{n-1}\left(\mathbb{E}_{\theta^0}\left[Y_i Y_{i+1}^T \mid X\right] R_y^{-1} - \mathbb{E}_{\theta^0}\left[Y_i Y_i^T \mid X\right] A^T R_y^{-1}\right).$$

Если приравняем к нулю и произведём сокращения, то получим, что оптимальная

$$A = \left(\sum_{i=1}^{n-1} P_{i,i+1}\right)\left(\sum_{i=1}^{n-1} P_{i,i}\right)^{-1}, \quad P_{i,i} = \mathbb{E}_{\theta^0}\left[Y_i Y_i^T \mid X\right], \quad P_{i,i+1} = \mathbb{E}_{\theta^0}\left[Y_i Y_{i+1}^T \mid X\right],$$

тут мы впервые ввели для краткости новое обозначение $P_{i,i}$ и $P_{i,i+1}$.

Оптимизируем по R_y. Здесь нам понадобится знакомая по гауссовскому ММП формула (3) для

$$\partial_C \ln det(C) = (C^T)^{-1}.$$

В итоге получим

$$\partial_{R_y} Q = -\frac{n}{2} R_y^{-1} + \frac{1}{2}\sum_{i=1}^{n-1} \left(R_y^{-1} P_{i+1,i+1} R_y^{-1} - R_y^{-1} A P_{i,i+1} R_y^{-1} - R_y^{-1} P_{i+1,i} A^T R_y^{-1} + R_y^{-1} A P_{i,i} A^T R_y^{-1} \right)$$

и, приравняв к нулю и сократив лишнее, придём к результату

$$R_y = \frac{1}{n}\sum_{i=1}^{n-1} \left(-A P_{i,i+1} - P_{i+1,i} A^T + P_{i+1,i+1} + A P_{i,i} A^T \right).$$

Оптимизируем по B и R_x. Здесь всё происходит абсолютно по тем же принципам, но возникнут произведения X и Y. Выпишем просто ответ:

$$B = \left(\sum_{i=1}^{n} \widetilde{xy}_i \right) \left(\sum_{i=1}^{n} P_{i,i} \right)^{-1}, \quad R_x = \frac{1}{n}\sum_{i=1}^{n} \left(X_i X_i^T + B P_{i,i} B^T - B\widetilde{xy}_i^T - \widetilde{xy}_i B^T \right),$$

где мы по аналогии с $P_{i,i}$ ввели

$$\widetilde{xy}_i = \mathbb{E}_{\theta^0} \left[X_i Y_i^T \mid X \right].$$

В итоге. Мы получили формулы для вычисления М-шага, но проблема в том, что все они зависят от матожиданий, которые мы пока не вычисляли. На первый взгляд это кажется сложной задачей, потому что самый прямой подход требует вычисления условного распределения по аналогии с тем, как мы делали в случае смесей. Но, возможно, вам кажется что-то знакомым...

5.3 Алгоритм фильтрации Калмана

Конечно, ровно такие матожидания вычислялись в фильтре Калмана! Конкретнее, в алгоритме RTSS (сглаживание). Если мы будем использовать алгоритм Калмана и RTSS, на выходе мы получим ровно нужные величины и завершим конструкцию ЕМ-алгоритма. Его можно будет пробовать запустить.

Вспомним модель как раньше, но с известными параметрами $A, B, R_x, R_y, \Lambda, \xi$ (они нам даны в θ_0, такова конструкция ЕМ-алгоритма).

$$Y_{t+1} = A y_t + U_{t+1},\ U_{t+1} \sim \mathcal{N}(0, R_y),$$
$$X_{t+1} = B y_{t+1} + W_{t+1},\ W_{t+1} \sim \mathcal{N}(0, R_x).$$

5.3.1 Финальный алгоритм

Фильтрация. Итак, алгоритм фильтрации Калмана. Задаём из параметров θ_0(по конструкции EM-алгоритма они даны) $\hat{Y}_1 = \xi, P_1 = \Lambda$, а далее в цикле по t

$$\tilde{Y}_t = A\hat{Y}_{t-1} \ \text{(априорный прогноз)},$$
$$\tilde{P}_t = AP_{t-1}A^T + R_y \ \text{(ошибка априорного прогноза)},$$
$$K_t = \tilde{P}_t B^T (B\tilde{P}_t B^T + R_x)^{-1} \ \text{(фильтр)}$$
$$\hat{Y}_t = \tilde{Y}_t + K_t(X_t - B\tilde{Y}_t) \ \text{(апостериорный прогноз)},$$
$$P_t = \tilde{P}_t - K_t B\tilde{P}_t \ \text{(ошибка апостериорного прогноза)}.$$

Сглаживание. Фильтрации нам мало, потому что она вычисляет только

$$\hat{Y}_t = \mathbb{E}\left[Y_t \mid X_1, .., X_t\right],$$

то есть, условное матожидание при условии моментов времени *до* t. Чтобы уточнить старые прогнозы фильтрации применяют алгоритм сглаживания (RTSS), а ещё дополнительно нам нужно вычислять ковариацию значений с лагом (детали опустим, формулу нужно получать из RTSS). По инструкции: задать $\overline{Y}_T = \hat{Y}_T$, $\overline{P}_T = P_T$, $V_{T-1,T} = 0$, а далее в цикле $t = T-1, .., 1$ выполнять

$$S_t = P_t A^T (\widetilde{P}_{t+1})^{-1} \ \text{(вычисление сглаживающего фильтра)},$$
$$\overline{Y}_t = \hat{Y}_t + S_t(\overline{Y}_{t+1} - \widetilde{Y}_{t+1}) \ \text{(сглаживание)},$$
$$\overline{P}_t = P_t + S_t(\overline{P}_{t+1} - \widetilde{P}_{t+1})S_t^T \ \text{(ошибка сглаживания)},$$
$$V_{t,t+1} = P_t S_t^T + K_{T-1}(V_{t+1,t+2} - AP_t)S_{t-1}^T, \ \text{(лаг-1 ковариация, для } t \leq T-2)$$

Остаётся только понять, как алгоритм фильтрации помогает достроить EM-алгоритм. Если вернуться немного назад, то интересующие матожидания можно найти в расчётах:

$$P_{i,i} = \overline{P}_i + \hat{Y}_i\hat{Y}_i^T, \quad P_{i.i+1} = V_{i,i+1} + \hat{Y}_i\hat{Y}_{i+1}^T,$$

а кросс-члены в силу известных наблюдений X задаются как

$$\widetilde{yx}_i = \hat{Y}_i X_i^T.$$

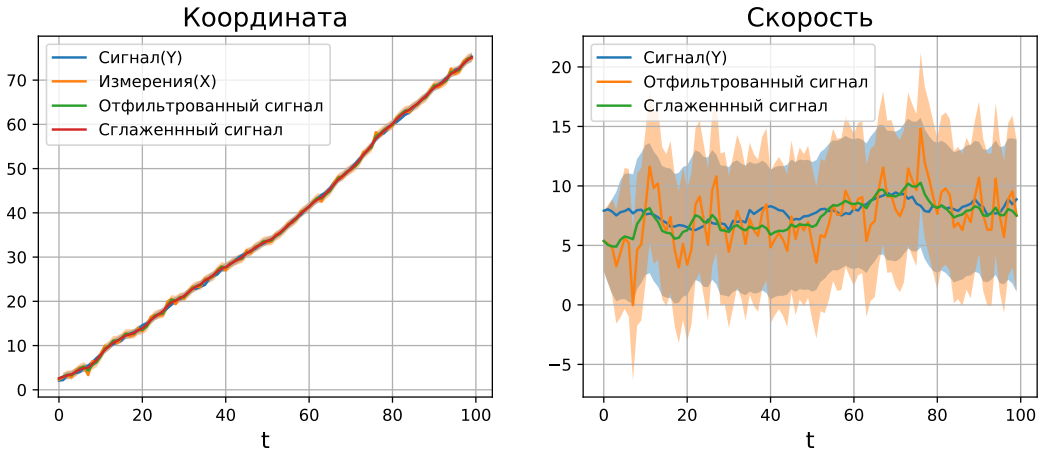
Итоговый EM-алгоритм выглядит теперь так:

1. Инициализировать $\theta^0 = \{A, B, R_x, R_y, \xi, \Lambda\}$;
2. Пока не остановимся:
   (a) E-шаг: Вычислить матожидания для параметров θ^0, используя алгоритм фильтрации и сглаживания;
   (b) M-шаг: вычислить новые параметры θ по выведенным формулам.

(лаг1: 0, а потом $V_{t-1,t} = P_{t-1}S_{t-1}^T + K_{T-1}(V_{t,t+1} - AP_t)S_{t-1}^T$)

5.3.2 Проблемы

...такой алгоритм может не сойтись. Если оценивать всё, то в системе очень много степеней свободы. Заметьте, к примеру, что матрицы A и B могут друг друга компенсировать и друг под друга подгоняться. Например, на шаге 10 матрица A диагональная с 10 на диагонали, матрица B – с числом 1/10; если они поменяются местами и правильно пересчитаются ковариационные матрицы R_X, R_y, то правдоподобие не изменится. На практике это соответствует тому, что наблюдатель Y меряет в килограммах интересующее значение, а наблюдающий X меряет значение в тоннах; а на более поздней итерации наоборот. Чтобы избежать этой проблемы достаточно зафиксировать одну из матриц, например, единичной. К счастью, в любой задаче для фильтра у вас уже есть априорная информация, позволяющая оценивать только часть неизвестных параметров. В этом случае алгоритм уже сойдётся и так будет выглядеть его результат после 100 итераций (доверительный интервал уровня 0.975 построен с использованием оценки апостериорной ошибки):



Если вдруг вы линейный скептик (и не верите в линейные модели), то в последних двух главах вы смогли увидеть тот самый случай, когда они способны на по-настоящему реальные вещи.

Литература

[1] T. W. Anderson and D. A. Darling. Asymptotic Theory of Certain "Goodness of Fit"Criteria Based on Stochastic Processes. *The Annals of Mathematical Statistics*, 23(2):193 – 212, 1952.

[2] Peter C Austin, Muhammad M Mamdani, David N Juurlink, and Janet E Hux. Testing multiple statistical hypotheses resulted in spurious associations: a study of astrological signs and health. *J Clin Epidemiol*, 59(9):964–969, July 2006.

[3] Erika Cule, Paolo Vineis, and Maria De Iorio. Significance testing in ridge regression for genetic data. *BMC Bioinformatics*, 12(1):372, Sep 2011.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[5] B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

[6] R. E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, 82(1):35–45, 03 1960.

[7] S.J. Levitt, S.D. Dubner. *Freakonomics*. NY: Harper Trophy, 2006.

[8] Richard Lockhart, Jonathan Taylor, Ryan J. Tibshirani, and Robert Tibshirani. A significance test for the lasso. *The Annals of Statistics*, 42(2):413–468, 2014.

[9] J. Scott Long and Laurie H. Ervin. Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, 54(3):217–224, 2000.

[10] Herbert E Rauch, F Tung, and Charlotte T Striebel. Maximum likelihood estimates of linear dynamic systems. *AIAA journal*, 3(8):1445–1450, 1965.

[11] O. A. Stepanov. Kalman filtering: Past and present. an outlook from russia. (on the occasion of the 80th birthday of rudolf emil kalman). *Gyroscopy and Navigation*, 2(2):99–110, Apr 2011.

[12] R. L. Stratonovich. Conditional markov processes. *Theory of Probability & Its Applications*, 5(2):156–178, 1960.

[13] Halbert White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838, 1980.

[14] Платонов Е.Н. Горяинова Е.Р., Панков А.Р. *Прикладные методы анализа статистических данных.* Изд. дом Высшей школы экономики, 2012.

[15] Ю.М. Кельберт, М.Я. Сухов. *Вероятность и статистика в примерах и задачах. Т.3: теория информации и кодирования.* М.: МЦНМО, 2013.

[16] Р.Л. Стратонович. *Условные марковские процессы и их применение к теории оптимального управления.* Московский государственный университет, 1966.

[17] А.Н. Ширяев. *Основы стохастической финансовой математики.* МЦНМО, 2016.