ЛЕКЦИЯ 4

ЕМ-алгоритм: оценка фильтров

Сейчас мы рассмотрим ещё одно приложение ЕМ-алгоритма, вошедшее в практику почти сразу после того, как ЕМ-алгоритм стал широко известен в 1977 году [3]. Речь пойдёт о фильтре Калмана [5], одном из центральных инструментов в обработке сигналов в робототехнике и различных приборах, включающих в себя датчики.

4.1 Задача фильтрации и извлечение сигнала

Представим себе, что у нас есть линейная модель

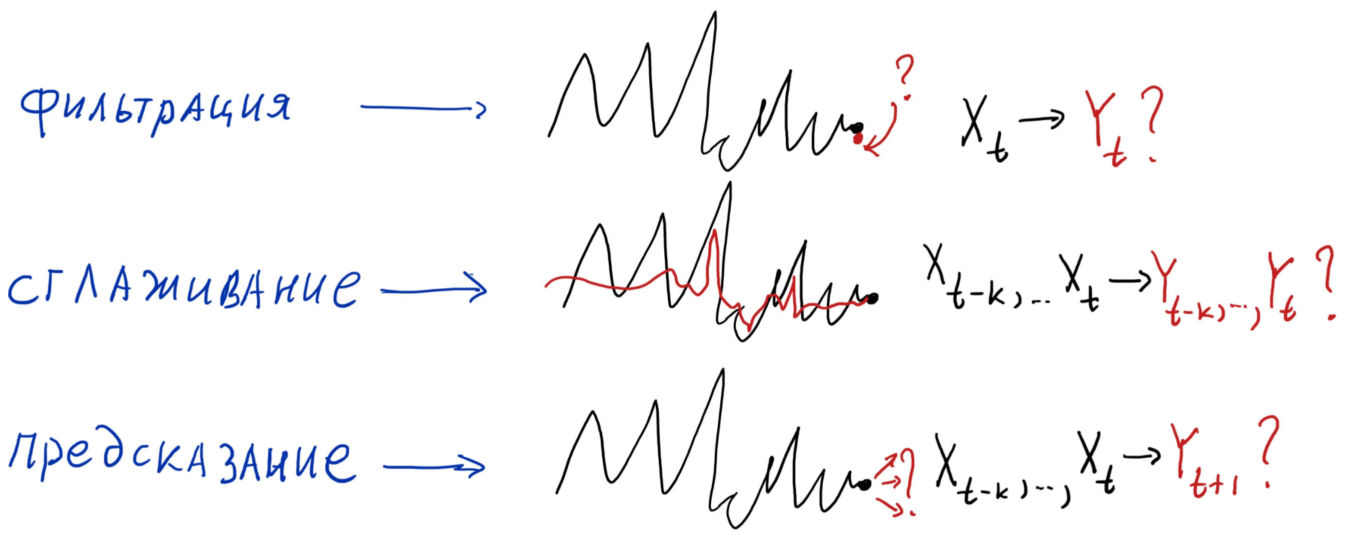
$$Y_{t+1} = AY_t + U_{t+1},$$

описывающая некоторую техническую систему (например, движение автомобиля или робота) с точностью до некоторого шума U_t. Такая модель оказывается вполне неплохой как минимум для небольших временных масштабов, а далее можно предполагать, что будут меняться матрица A или параметры шума. Это реальный мир, но мы его наблюдаем через какие-то приборы (например, GPS-маячок), которые сигнал некоторым образом искажают и тоже добавляют свой шум:

$$X_{t+1} = BY_{t+1} + W_{t+1}.$$

Мы наблюдаем только набор X_t и хотели бы по максимуму избавиться от шума и попробовать оценить настоящий сигнал Y_t. При этом в зависимости от конкретного приложения нас может интересовать это оценивание в разном контексте:

1. *Задача фильтрации.* При известном X_t на ходу оценивать Y_t.
2. *Задача сглаживания.* При известном $X_{t-n}, .., X_t$ оценить $Y_{t-n}, .., Y_t$.
3. *Задача предсказания.* При известном $X_{t-n}, .., X_t$ оценить Y_{t+1}.



Оказывается, что в данных условиях универсальным решением является алгоритм, который носит название *фильтра Калмана* и был впервые описан Калманом в 1960 году [5], а после распространения EM-алгоритма появились процедуры оценивания параметров этой и более общей нелинейной модели.

4.2 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 мы для сокращения нотаций предполагаем из $\mathbb{R}^{n\times n}$. 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 являются матрицы A, B, ковариационные матрицы векторов шума R_x, R_y, а также параметры стартового состояния $Y_0 \sim \mathcal{N}(\mu_0, R_0)$. 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

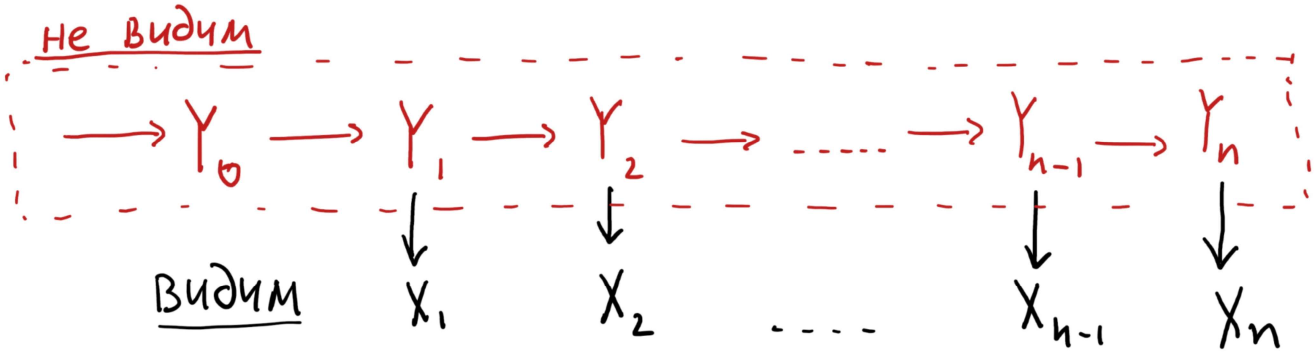
$$Q(\theta^0|\theta) := \mathbb{E}_{\theta^0}\left[\ln p_\theta(X, Y) \mid X\right]$$

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

$$\ln p_\theta(Y) = \ln p_\theta(Y_0) + \sum_{i=1}^{n-1} p_\theta(Y_{i+1}|Y_i).$$

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. С последней работать непросто, так как невозможно одновременно оценить и матожидание, и ковариацию; поэтому мы для простоты зафиксируем $R_0 = \Sigma_0$ (популярно просто диагональной матрицей), а $\mu_0 = X_1$.

Если мы воспользуемся теперь явным видом распределений и отбросим стартовый член с Y_1 в одну общую константу, получим

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

А после раскрытия скобок

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

в котором 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 $\theta = \{A, B, R_x, R_y\}$. Заметьте, что матожидания берутся по модели данных с парамтерами θ^0, поэтому мы можем вносить производные по θ под знак матожидания.

4.3 Оценка параметров: М-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)

$$\partial_D \left(b^T D^T a\right) = \partial_D \left(a^T D b\right) = b a^T.$$

и более сложная фомула (2)

$$\partial_D \left(b^T D^T C D d\right) = d b^T D^T C + b d^T D^T C^T.$$

Применим, отбросив не зависящие от A члены:

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

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

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

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

Оптимизируем по R_y. Здесь нам понадобится формула (3)

$$\partial_C (a^T C^{-1} c) = -C^{-1} a c^T C^{-1},$$

а также знакомая по гауссовскому ММП формула (4) для

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

В итоге получим

$$\partial_{R_y} Q = -\frac{n}{2} R_y^{-1} + \frac{1}{2} \sum_{i=0}^{n-1} \left(R_y^{-1} P_{i+1,i+1} R_y^{-1} - R_y^{-1} A P_{i,i+1} R_y^{-1} - R_y^{-1} P_{i+1,i} A^T R_y^{-1} + R_y^{-1} A P_{i,i} A^T R_y^{-1} \right)$$

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

$$R_y = \frac{1}{n} \sum_{i=0}^{n-1} \left(-AP_{i,i+1} - P_{i+1,i}A^T + P_{i+1,i+1} + AP_{i,i}A^T \right).$$

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

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

где мы по аналогии с $P_{i,i}$ ввели

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

В итоге. Мы получили формулы для вычисления M-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 выход есть.

4.4 Алгоритм фильтрации Калмана

Калман рассматривал задачу фильтрации: как предсказать Y_t по известной истории и X_t, если параметры A, B, R_x, R_y уже даны. Это ровно наш случай: матожидания берутся в предположении вероятностной модели с параметрами $\theta^0 = \{A^0, B^0, R_x^0, R_y^0\}$, то есть, с известными в рамках EM-алгоритма. Если мы научимся вычислять эти матожидания, мы завершим конструкцию EM-алгоритма.

Почему вообще может быть интересна такая постановка, то есть, без оценки параметров модели? Она просто другая: в главе выше мы занимались оценкой модели. Сейчас в предположении известной модели мы пытаемся построить *наилучший* (в смысле минимальной квадратичной ошибки) способ оценки целевого сигнала в условиях имеющегося модельного предположения. Как раз здесь и возникает задача фильтрации, задача сглаживания и задача предсказания. Положим модель как раньше, но с известными параметрами(мы будем для краткости опускать верхний индекс 0 у параметров A, B, R_x, R_y).

$$Y_{t+1} = Ay_t + U_{t+1},\ U_{t+1} \sim \mathcal{N}(0, R_y),$$
$$X_{t+1} = By_{t+1} + W_{t+1},\ W_{t+1} \sim \mathcal{N}(0, R_x).$$

Как зная X_t оценить Y_t? Идея в следующем:

1. Вычислим некоторым образом, используя историческую информацию в момент $t-1$, априорную оценку $\hat{Y}_t'$ (пока предположим, что дана свыше);

2. Посмотрим, какое предсказание наблюдения X_t получается, если реальный сигнал был бы $\hat{Y}_t'$, а затем попробуем скорректировать предсказание с помощью отклонения $X_t - B\hat{Y}_t'$:
$$\hat{Y}_t = \hat{Y}_t' + K_t\left(X_t - B\hat{Y}_t'\right).$$

Здесь K_t – это матрица, называемая часто матрицей фильтрации или *Kalman gain.*

4.4.1 Вычисление фильтра

Калман предложил на каждом шаге t выбирать эту матрицу так, чтобы минимизировать квадрат нормы ошибок

$$\mathrm{tr}E_t = \mathrm{tr}\mathbb{E}_\theta\left[(Y_t - \hat{Y}_t)(Y_t - \hat{Y}_t)^T \mid X\right],$$

отсюда получается оптимальность фильтра. Здесь важно сделать оговорку: в силу того, что
$$\mathbb{E}_\theta\left[(Y_t - \hat{Y}_t)(Y_t - \hat{Y}_t)^T\right] = \mathbb{E}_\theta\left[\mathbb{E}_\theta\left[(Y_t - \hat{Y}_t)(Y_t - \hat{Y}_t)^T \mid X\right]\right],$$
где внешнее матожидание берётся только по наблюдаемым значениям и это распределение не зависит от способа прогноза, задачи

$$\min_{K_t} \mathrm{tr}E_t \quad \text{и} \quad \min_{K_t} \mathrm{tr}\mathbb{E}_\theta\left[(Y_t - \hat{Y}_t)(Y_t - \hat{Y}_t)^T\right]$$

имеют одно и то же решение – то есть, будущие сигналы во времена $t+1, t+2, ..$ знать не обязательно.

Чтобы примерно понять, как вычислить K_t, перепишем выражение для ковариации ошибок, раскрыв скорректированное предсказание согласно пункту 2:

$$E_t = \mathbb{E}_\theta\left[\left((I - K_tB)(Y_t - \hat{Y}_t') - K_tW_t\right)\left((I - K_tB)(Y_t - \hat{Y}_t') - K_tW_t\right)^T \mid X\right] =$$
$$= (I - K_tB)\mathbb{E}_\theta\left[(Y_t - \hat{Y}_t')(Y_t - \hat{Y}_t')^T \mid X\right](I - K_tB)^T + K_tR_xK_t^T.$$

Посередине мы видим ошибку апирорного предсказания

$$E_t' = \mathbb{E}_\theta\left[(Y_t - \hat{Y}_t')(Y_t - \hat{Y}_t')^T \mid X\right].$$

Если бы мы её знали, то, мы могли бы уже вычислить решение задачи

$$\min_{K_t} \mathrm{tr}E_t = \min_{K_t} \mathrm{tr}(I - K_tB)E_t'(I - K_tB)^T + K_tR_xK_t^T.$$

Используя те же формулы дифференцирования и ещё несколько новых формул для производной следа, можно через некоторое время (опустим технические детали) получить

$$K_t = E_t' B^T (R_x + B E_t' B^T)^{-1}$$

Более того, если подставим в выражение для E_t, то увидим, как связаны апостериорная (после коррекции) ошибка E_t и априорная E_t':

$$E_t = E_{t'}(I - B^T K_t^T) = (I - K_t B) E_t'.$$

4.4.2 Априорное предсказание и априорная ошибка

Мы на шаге t, нам известна модель и предыдущие (мы надеемся) наилучшие предсказания, как можно построить априорную оценку? Наилучший прогноз – это условное матожидание при условии того, что известно. Так из динамики мы уже можем прикинуть ответ

$$\hat{Y}_t = A\hat{Y}_{t-1}.$$

Остаётся нерешённой проблема вычисления ошибок, нам нужно как-то уметь оценивать априорную ошибку E_t'. Давайте воспользуемся известной динамикой и заметим, что

$$Y_t - \hat{Y}_t' = (AY_{t-1} + U_t) - A\hat{Y}_{t-1} = A(Y_{t-1} - \hat{Y}_{t-1}) + U_t.$$

То есть, она зависит от предыдущей апостериорной ошибки:

$$E_t' = AE_{t-1}A^T + R_y.$$

4.4.3 Алгоритм фильтрации

Подытожим, что мы знаем про фильтрацию.

1. Инициализация: $\hat{Y}_0' = X_1, E_0 = \sigma_0^2 I$;

2. Проходим итерации до $t = n$:

$$\begin{aligned}
&\hat{Y}_t' = A\hat{Y}_{t-1} && \text{(вычисляем априорное предсказание)},\\
&E_t' = AE_{t-1}A^T + R_y && \text{(вычисляем априорную ошибку)},\\
&K_t = E_t' B^T (R_x + B E_t' B^T)^{-1} && \text{(вычисляем фильтрацию)},\\
&\hat{Y}_t = \hat{Y}_t' + K_t(X_t - B\hat{Y}_t') && \text{(вычисляем коррекцию)},\\
&E_t = (I - K_t B) E_t' && \text{(вычисляем апостериорную ошибку)}.
\end{aligned}$$

Априорное предсказание решает задачу предсказания, коррекция решает задачу фильтрации, а задача сглаживания решается применением этого алгоритма к набору наблюдений $X_1, .., X_n$. Фильтр Калмана универсально адресует все три задачи и по конструкции делает это наилучшим образом в смысле квадратичной ошибки.

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

$$P_{i,i} = E_i + \hat{Y}_i \hat{Y}_i^T, \quad P_{i.i-1} = AE_{i-1} + \hat{Y}_i \hat{Y}_{i-1}^T,$$

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 $\theta^0 = \{A, B, R_x, R_y\}$;

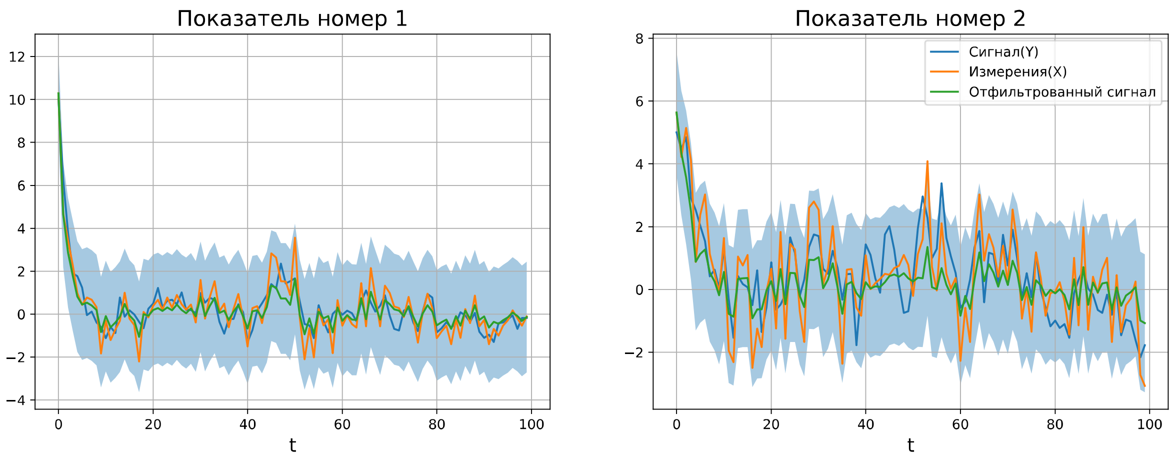
2. Пока не остановимся:

   (a) E-шаг: Вычислить матожидания для параметров θ^0, используя алгоритм фильтрации(гл.4.4);

   (b) M-шаг: вычислить новые параметры θ, используя результат алгоритма фильтрации(гл.4.3).

4.4.4 Проблемы

...такой алгоритм не сойдётся. Проблема в том, как мы выбрали параметры модели. Заметьте, 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, 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. Обычно просто делают $B = Id$ и тогда алгоритм уже сойдётся и так будет выглядеть его результат после 100 итераций (доверительный интервал уровня 0.975 построен с использованием оценки апостериорной ошибки):


Показатель номер 1
Показатель номер 2
Сигнал(Y)
Измерения(X)
Отфильтрованный сигнал
t
t

Литература

[1] 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.

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

[3] 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.

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

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

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

[7] 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.

[8] 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.

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

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

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