CS4070: EXERCISE 1: BAYESIAN LINEAR REGRESSION AND KALMAN FILTERING

Hand in before December 9, 12.00

In the following we use "Bayesian notation" throughout.

1. Bayesian updating for linear regression

Suppose we have observations $y_1, \ldots, y_n$ satisfying a linear regression model

$$y_i = \theta_1 + \theta_2 t_i + \varepsilon_i \qquad \varepsilon_i \overset{\text{ind}}{\sim} N\left(0, \sigma^2\right).$$

The times $t_1 < t_2 < \ldots$ are the observation times. We assume for simplicity that σ^2 is known. If we define

$$H_i = \begin{bmatrix} 1 & t_i \end{bmatrix} \qquad \theta = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix},$$

then we can can write

$$y_i \sim N\left(H_i\theta, \sigma^2\right).$$

Define $y = \begin{bmatrix} y_1 & \ldots & y_n \end{bmatrix}'$. The likelihood is given by

$$L(\theta \mid y) = p(y \mid \theta) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{1}{2\sigma^2}(y_i - H_i\theta)^2\right).$$

That is,

$$p(y \mid \theta) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2}(y - H\theta)'(\sigma^2 I_n)^{-1}(y - H\theta)\right),$$

where

$$H = \begin{bmatrix} H_1 \\ \vdots \\ H_n \end{bmatrix} = \begin{bmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_n \end{bmatrix}.$$

Clearly, $y \mid \theta \sim N_n(H\theta, \sigma^2 I_n)$. We take $\theta \sim N_2(m_0, P_0)$ a priori.

Exercise 1. Show that $\theta \mid y \sim N_2(\nu, C)$, where

$$C^{-1} = H'\sigma^{-2}H + P_0^{-1}$$

and

$$\nu = C\left(H'\sigma^{-2}y + P_0^{-1}m_0\right).$$

That is, both the prior and posterior distribution are normal. Put differently, the chosen prior is conjugate for the given statistical model.

Date: November 22, 2019.

Bayesian updating refers to the following observation: if we let $y_{1:k} = \begin{bmatrix} y_1 & \cdots & y_k \end{bmatrix}'$ then

$$\begin{aligned} p(\theta \mid y_{1:k}) &\propto p(y_{1:k} \mid \theta) p(\theta) \\ &= p(y_{1:k-1} \mid \theta) p(y_k \mid \theta) p(\theta) \\ &\propto p(y_k \mid \theta) p(\theta \mid y_{1:k-1}). \end{aligned}$$

The equality on the second line follows from $y_{1:k-1}$ and y_k being independent, conditional on θ. Therefore, if we wish to find the posterior after k observations, we can obtain it by considering only the k-th observation coming in with prior distribution for θ equal to the posterior of θ based on the first $k-1$ observations.

Suppose that $\theta \mid y_{1:k} \sim N(m_k, P_k)$. Then

$$P_k^{-1} = H_k' \sigma^{-2} H_k + P_{k-1}^{-1} \tag{1}$$

and

$$m_k = P_k \left(H_k' \sigma^{-2} y_k + P_{k-1}^{-1} m_{k-1} \right).$$

The case $k = 1$ corresponds to question 1.

Exercise 2. Use the Woodbury matrix identity (https://en.wikipedia.org/wiki/Woodbury_matrix_identity) to show that

$$P_k = P_{k-1} - P_{k-1} H_k' \left(H_k P_{k-1} H_k' + \sigma^2 \right)^{-1} H_k P_{k-1}.$$

Why is it numerically advantageous to use this formula for updating $\{P_k\}$ over inverting the right-hand-side in equation (1)?

Exercise 3. Download the data in `SalaryData.csv`, which contains for a number of employees the number of years of working experience (t) and salary (y). Assume y depends on t polynomially with degree 3, i.e.

$$y_i = \theta_1 + \theta_2 t_i + \theta_3 t_i^2 + \theta_4 t_i^3 + \varepsilon_i.$$

(1) Implement an algorithm that sequentially computes the posterior distribution. That is, at each iteration, one row in the csv-file containing the data is used as "incoming data".

Assume apriori that $\theta = (\theta_1, \theta_2, \theta_3, \theta_4) \sim N((40000, 0, 0, 0), \sigma_0^2 I)$ with $\sigma_0^2 = 100$. For the measurement standard deviation assume $\sigma = 250$ is known.

(2) Report the posterior mean and covariance matrix of $\theta = (\theta_1, \ldots, \theta_4)$ based on the first 10 observations. Also report these quantities based on the full dataset.

(3) Present a figure with the fitted curve when using all observations. Superimpose the observed data.

(4) Include your code as an appendix. Ensure this code is readable and sufficiently well documented.

There is nothing special about the chosen form of H_k, the just derived updating formulas hold generally under the assumption that $y_1, \ldots, y_n$ are independent (conditional on θ) with $y_k \mid \theta \sim N\left(H_k \theta, \sigma^2\right)$. Now let's assume the parameter θ is not constant, but in fact a signal that evolves over time. Say we have

$$\theta_k = A\theta_{k-1} + q_{k-1} \qquad q_{k-1} \sim N\left(0, Q\right).$$

So in total we have the model

$$
\begin{aligned}
y_k &= H_k\theta_k + \varepsilon_k \qquad \text{observation model} \\
\theta_k &= A\theta_{k-1} + q_{k-1} \qquad \text{signal}
\end{aligned}
$$

This is an example of a linear *state-space model*. We could for instance have that $\theta_k \in \mathbb{R}^2$ and $H_k = \begin{bmatrix} 1 & 0 \end{bmatrix}$. This corresponds to only observing the first component of the signal with noise. We aim to sample from $\theta_k \mid y_{1:k}$. This is known as the *filtering problem*. If we can do this, then we are able to reconstruct/estimate not only the first component of the signal, but the second component as well!

Suppose for simplicity that $\{H_k\}_{k=1}^n$, A, Q and σ^2 are known. At time 0, before any observation has been obtained, we assume $\theta_0 \sim N(m_0, P_0)$ (just as in the previous section; this is the prior). The *Kalman* filter gives the formulas for updating $\theta_{k-1} \mid y_{1:k-1}$ to $\theta_k \mid y_{1:k}$. It consists of two steps:

(1) The *prediction step*. We have

$$\theta_k \mid y_{1:k-1} \sim N\left(m_k^-, P_k^-\right)$$

with

$$
\begin{aligned}
m_k^- &= Am_{k-1} \\
P_k^- &= AP_{k-1}A' + Q
\end{aligned}
\tag{2}
$$

(2) The *update step*. Here, we use $p(\theta_k \mid y_{1:k-1})$ as a prior for the incoming observation $y_k \sim N\left(H_k\theta_k, \sigma^2 I\right)$. As previously derived we have $\theta_k \mid y_{1:k} \sim N(m_k, P_k)$ with

$$P_k = P_k^- - P_k^- H_k' \left(H_k P_k^- H_k' + \sigma^2\right)^{-1} H_k P_k^-.$$

and

$$m_k = P_k^- \left(H_k'\sigma^{-2}y_k + \left(P_k^-\right)^{-1} m_k^-\right).$$

Exercise 4. Verify the formulas in (2) for the prediction step of the Kalman filter. *Hints:*

(1) Note that the distribution of $\theta_k \mid y_{1:k-1}$ can be obtained as the marginal distribution of $(\theta_k, \theta_{k-1}) \mid y_{1:k-1}$.

(2) Explain why

$$p(\theta_k, \theta_{k-1} \mid y_{1:k-1}) = p(\theta_k \mid \theta_{k-1})p(\theta_{k-1} \mid y_{1:k-1}).$$

(3) Apply lemma 1 below to deduce that the joint distribution of $(\theta_k, \theta_{k-1}) \mid y_{1:k-1}$ is multivariate normal with the given parameters.

Lemma 1. If the random vectors $X \in \mathbb{R}^n$ and $Y \in \mathbb{R}^m$ satisfy

$$
\begin{aligned}
X &\sim N(m, P) \\
Y \mid X &\sim N(Hx + u, R)
\end{aligned}
$$

then

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim N\left(\begin{bmatrix} m \\ Hm + u \end{bmatrix}, \begin{bmatrix} P & PH' \\ HP & HPH' + R \end{bmatrix}\right).$$