STA642 hw4

Gongjinghao Cheng; gc169

# Q3

## (4.3.1)

Suppose $\theta_{t-1}|D_{t-1} \sim MVN(m_{t-1}, C_{t-1})$, since $\theta_t = G_t\theta_{t-1} + w_t$, we know that $\theta_t|\theta_{t-1}$ is also a multivariate normal (with a linear transformation of G and translation of $w_t$), which is $MVN(a_t = G_t m_{t-1}, R_t = G_t C_{t-1} G_t' + W_t)$.

For $y_t|D_{t-1}$, with $\theta_{t-1}|D_{t-1}$, we can see that this is alse a linear transformation of a multivariate normal random variable, $(\theta_{t-1}|D_{t-1})$, thus $y_t|D_{t-1}, \theta_{t-1} \sim MVN(f_t = F_t' a_t, q_t = F_t' R_t F_t + v_t)$.

To derive the posterior $\theta_t|D_t$. First note that if we create a vector $\begin{bmatrix}\theta_t \\ y_t\end{bmatrix}|D_{t-1}$, and this is a multivariate normal distribution. Compute the covariance:

$$\begin{aligned}
&C(\theta_t, y_t) = E[(\theta_t - a_t)(y_t - f_t)'] \\
&= E[\theta_t y_t' - a_t y_t' - \theta_t f_t' + a_t f_t'] \\
&= E[\theta_t (F_t'\theta_t)^T] - a_t (F_t' a_t)' \\
&= E[\theta_t \theta_t'] F_t - a_t a_t' F_t \\
&= R_t F_t
\end{aligned}$$

Thus, $\begin{bmatrix}\theta_t \\ y_t\end{bmatrix}|D_{t-1} \sim MVN(\begin{bmatrix}a_t \\ f_t\end{bmatrix}, \begin{bmatrix}R_t & R_t F_t \\ F_t' R_t' & q_t\end{bmatrix})$.

Apply the conditional distribution formula of multivariate normal distribution, we get$\theta_t|y_t, D_t \sim N(a_t + R_t F_t q_t^{-1}(y_t - f_t), R_t - R_t F_t q_t^{-1} F_t^T R_t^T) = N(m_t, C_t)$ as in 4.8.

## (4.3.2)

Derive the posterior, let r be the precision, 1/v:

$$\begin{aligned}
&p(\theta_t|D_t) \\
&\propto \int p(\theta_t, D_t, r) dr \\
&\propto \int p(\theta_t|D_t, r) p(r|D_t) dr \\
&\propto \int NVM(\theta_t|m_t, \frac{1}{r s_t} C_t) Ga(r|n_t/2, n_t s_t/2) dr
\end{aligned}$$

Note that the marginal distribution will be multivariate T distribution by definition. Continue derivation:

$$\int NVM(\theta_t|m_t, \frac{1}{rs_t}C_t)Ga(r|n_t/2, n_t s_t/2)dr$$
$$\propto \int r^{1/2}exp(-\frac{1}{2}(\theta_t - m_t)^T rs_t C_t^{-1}(\theta_t - m_t))r^{n_t/2-1}exp(-\frac{1}{2}n_t s_t r)dr$$
$$\propto \frac{\Gamma((n_t + P)/2)}{n_t^{\frac{n_t+p}{2}}(\frac{1}{n_t}(\theta_t - m_t)^T C_t^{-1}(\theta_t - m_t) + 1)^{\frac{n_t+p}{2}}} \int Ga(r|\frac{n_t + p}{2}, 1/2[s_t((\theta_t - m_t)^T C_t^{-1}(\theta_t - m_t) + n_t)])dr$$
$$\propto (1 + \frac{1}{n_t}(\theta_t - m_t)^T C_t^{-1}(\theta_t - m_t))^{-\frac{n_t+p}{2}}$$

This is the kernel of $T_{n_t}(m_t, C_t)$.

## Q4

### (a)

Consider "Kalman filterïn this special case, we plug in $G_t$ and $W_t$. We get:

$$\theta_t|D_{t-1} \sim N(a_t, R_t)$$
$$a_t = G_t m_{t-1} = m_{t-1}, R_t = (1+\epsilon)C_{t-1} = \frac{C_{t-1}}{\delta}$$
$$y_t|D_{t-1} \sim N(f_t, q_t)$$
$$f_t = F_t' a_t = F_t' m_{t-1}, q_t = \frac{F_t' C_{t-1} F_t + \delta v_t}{\delta}$$
$$\theta_t|D_t \sim N(m_t, C_t)$$
$$A_t = \frac{R_t F_t}{q_t} = \frac{C_{t-1}F_t}{F_t' C_{t-1} F_t + v_t \delta}, m_t = a_t + A_t e_t = m_{t-1} + A_t e_t,$$
$$C_t = R_t - A_t A_t' q_t = R_t(I - \frac{F_t F_t' R_t'}{q_t}) = \frac{C_{t-1}}{\delta}(I - \frac{F_t F_t' C_{t-1}'}{F_t' C_{t-1} F_t + \delta v_t})$$

### (b)

First note that $m_t = m_{t-1} + A_t e_t$. The simplified structure does not need to evolve state matrix, therefore for the $m_t$ update it only depends on $m_{t-1}$ and the weighted forecast error. For update of $C_t$, it only depends on $C_{t-1}$ and regressor $F_t$ since the evolution noise $W_t$ is proportional to posterior variance of $\theta_t$.

For the impact of $\delta$. First note that large $\delta$ results in small $A_t$. With close examination, large $\delta$ will cause small difference between consecutive $m_{t-1}$ and $m_t$, and the posterior variance of $\theta$ will shrink.

### (c)

The simplified structure reduces computational cost by eliminating evolution of state matrix and assuming evolution noise proportional to posterior variance. This reduces the number of matrix multiplication in computation.

## Q5

### (a)

$$
\begin{aligned}
&C(\theta_t, \theta_{t-1}|D_{t-1}) \\
&= E[(\theta_t - a_t)(\theta_{t-1} - m_{t-1})^T] \\
&= E[\theta_t\theta_{t-1} - a_t\theta_{t-1}^T - \theta_{t}{}_{t-1}^T + a_t m_{t-1}^T] \\
&= G_t E[\theta_{t-1}\theta_{t-1}^T] - G_t m_{t-1} m_{t-1}^T \\
&= G_t V(\theta_{t-1|D_{t-1}}) \\
&= G_t C_{t-1}
\end{aligned}
$$

Hence, $C(\theta_{t-1}, \theta_t|D_{t-1}) = C_{t-1}G_t'$ follows.

### (b)

We know $\begin{bmatrix} \theta_t \\ \theta_{t-1} \end{bmatrix}$ is a 2p multivariate normal distribution. We solve this question with conditional distribution of multivariate normal distribution. First note that:

$$
\begin{bmatrix} \theta_t \\ \theta_{t-1} \end{bmatrix} \sim MVN(\begin{bmatrix} a_t \\ m_{t-1} \end{bmatrix}, \begin{bmatrix} R_t & G_t C_{t-1} \\ C_{t-1}G_t' & C_{t-1} \end{bmatrix})
$$

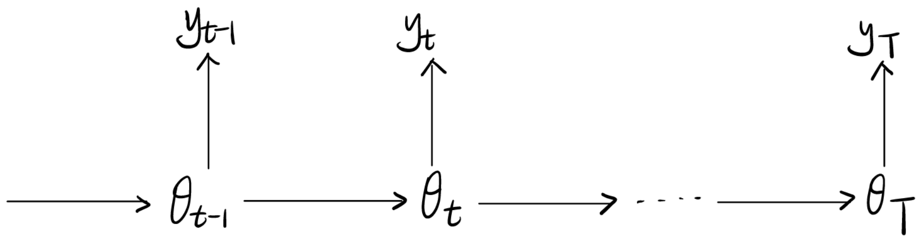
Then we can see that:

$$
\theta_{t-1}|\theta_t, D_{t-1} \sim MVN(m_{t-1} + C_{t-1}G_t' R_t^{-1}(\theta_t - a_t), C_{t-1} - C_{t-1}G_t' R_t^{-1} R_t R_t^{-1} G_t C_{t-1}')
$$

Equation 12, 13 follows as we take $B_{t-1} = C_{t-1}G_t' R_t^{-1}, m_{t-1}^* = m_{t-1} + B_{t-1}(\theta_t - a_t), C_{t-1}^* = C_{t-1} - B_{t-1}R_t B_{t-1}'$

### (c)

Note that with graphical model, $\theta_t$ does not depend on future $y_i$'s. Therefore, $p(\theta_{t-1}|\theta_t, D_n)$ is the same as $p(\theta_{t-1}|\theta_t, D_{t-1})$. This is intuitive since the Markov chain moves forward and the later observations should not influence the previous state space.




### (d)

This theory implies that:

$$p(\theta_{1:n}|D_n) = p(\theta_1|\theta_2, D_1)p(\theta_2|\theta_3, D_2)p(\theta_3|\theta_4, D_3)...p(\theta_n|D_n)$$

In this way we could state the retrospective distribution for the states $p(\theta_{1:n}|D_n)$.

### (e)

With this special case, we have:

$$\theta_{t-1}|\theta_t, D_{t-1} \sim MVN(m^*_{t-1}, C^*_{t-1})$$
$$m^*_{t-1} = m_{t-1} + C_{t-1}R_t^{-1}(\theta_t - m_{t-1}),$$
$$C^*_{t-1} = C_{t-1}(I - R_t^{-1}C_{t-1})$$
$$R_t = \frac{C_{t-1}}{\delta}$$

We can observe that the structure is simplified as we do not evolve the state matrix with G. The posterior variance, $R_t$ has simpler form with $W_t$ proportional to observational noise, further leads to simpler form of $C^*_{t-1}$.

Note that large $\delta$ results in smaller $R_t$, and smaller $C^*_{t-1}$ eventually; With G as identity, $m^*_{t-1}$ only differs from $m_{t-1}$ by weighted difference between current posterior mean, $(m_{t-1})$ and future $\theta$, $(\theta_t)$.

With this case, the computational cost is reduced since we abandon the needs of doing several matrix multiplication steps. More current information is depends on the priors (previous information).