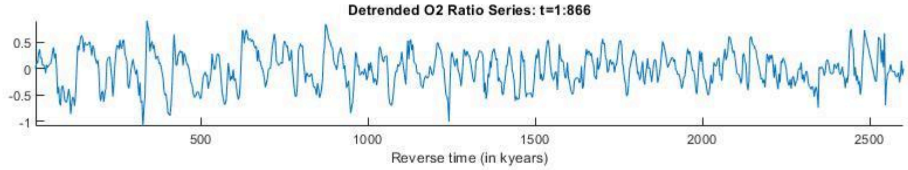
Midterm exam - Jae Hyun Lee(jl914)

Table of Contents

Question 1 - set up

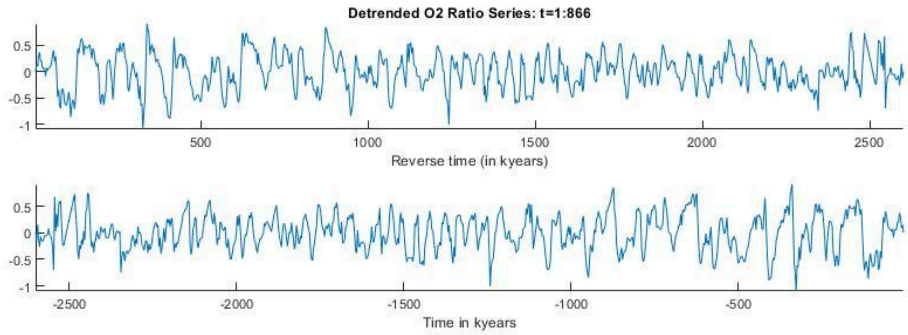
1. loading data

```
y=load('o2plusdetrended.txt'); ti=y(:,1); x=y(:,3); T=length(ti);
xa='box off; axis tight; xlabel(''Reverse time (in kyears)'');
 ylabel('''')';
figure(1); clf;
subplot(2,1,1); plot(ti,x);
title('Detrended O2 Ratio Series: t=1:866'); eval(xa);
```



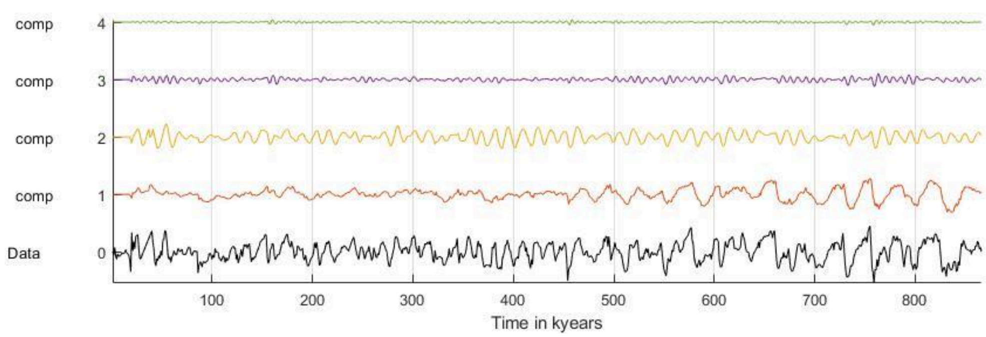
now make forward time as we are interested in forecasting from now ...

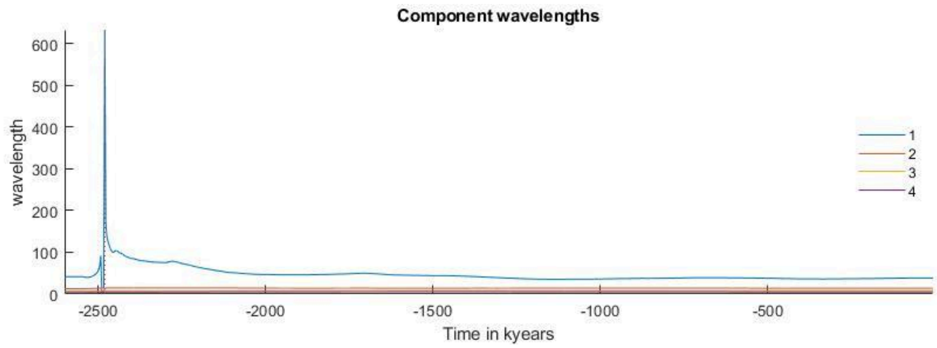
```
x=flipud(x); ti=-flipud(ti);
xa='box off; axis tight; xlabel(''Time in kyears''); ylabel('''')';
subplot(2,1,2); plot(ti,x); eval(xa);
```



1.(a)

```
p = 18;
del = [0.99, 0.975];
m0 = zeros(p,1); m0(1) = 1; n0 = 5; s0 = 0.02; C0 = eye(p)/2;

% Fit model
[m,C,n,s,e,mf,Cf,sf,nf,ef,qf] = tvar(x,p,del,m0,C0,s0,n0);

% Decomposition
[waves,mods,decomp,nr,nc] = tvar_decomp(x,m);

figure(1); clf ;
decomp_plot(x,decomp,4); eval(xa);

% plot plug-in estimate of wavelength
ic=min(4,size(waves,1));
figure(2); clf;
plot([ti,ti,ti,ti],waves(1:ic,:)');
eval(xa); ylabel('wavelength'); title('Component wavelengths');
hold on; plot([-2481,-2481], [0, max(waves(1,:))], ':r'); hold off;
legend( int2str((1:ic)'), 'location','east'); legend boxoff;
```





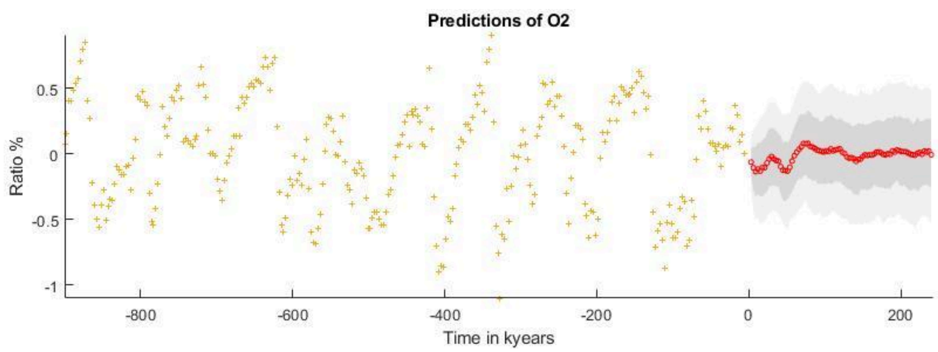
By investigating decomposition plot of data, we can find that first component dominates in amplitude in the process which has the largest wavelength and second component has also significant amplitude. Component3 and 4 have higher frequency but have much lower amplitude. The amplitude and frequency of components have changed especially in the first component. We can find that from -1500(indicated as 500)kyears ago, frequency significantly changed and have shorter wavelength which also indicated in figure2 as decrease of wavelength in first component. On the other hand, component3 and 4 are persistent. As in decomposition plot, component1 which has dominant moduli has the sharpest peak in wavelength indicated by the vertical dotted line drawn at the shapest peak and it decreases as time went. Corresponding to change of component1's frequency in figure1, wavelength of component 1 shows signficant change while the others keep constant. This dramatic change indicates the significant change in latent process. Thus, we cannot neglect the change of eigenstructure of evolution matrix and need to fit time varying model.

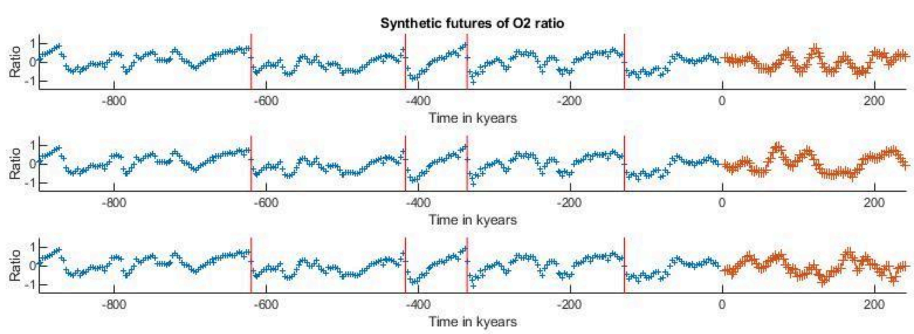
1.(b)

steps ahead and MC samples

```
rng(642)
k = 80;
I = 1000;
T = length(x);
mT=m(:,T); CT=squeeze(C(:,:,T)); sT=s(T); nT=n(T);
[ y,mu phi,v,probsns] = tvarforecast(x,k,I,p,del,mT,CT,sT,nT);

figure(3); clf;
nx=300;
pr = prctile(y',[10 25 50 75 90])';
ciplot(pr(:,1),pr(:,5),3*(1:k),[0.95 0.95 0.95]); hold on;
ciplot(pr(:,2),pr(:,4),3*(1:k),[0.85 0.85 0.85]);
scatter(3*(1:k),pr(:,3),10,'ro')
scatter(ti,x,10,'+'); hold on; ylim([min(x) max(x)]);
    %%errorbar(T+1:T+h,mean(y,2),1.5*std(y,0,2),'m*')
xa='box off; axis tight; xlabel(''Time in kyears''); ylabel('''')';
hold off; eval(xa); xlim(3*[-nx k+1]);
title(['Predictions of O2']);  ylabel(['Ratio %']);

figure(4); clf;
r=3; ir=randsample(I,r);
subplot(3,1,1);
scatter(ti,x,10,'+'); hold on
plot(3*(1:k),y(:,ir(1)),'+-');
```

```
plot([-620,-620], [-1.5,1.5], 'r');
plot([-417,-417], [-1.5,1.5], 'r');
plot([-336,-336], [-1.5,1.5], 'r');
plot([-129,-129], [-1.5,1.5], 'r');
hold off; eval(xa); xlim(3*[-nx k+1]); ylim([-1.5,1.5]);
title(['Synthetic futures of O2 ratio']);  ylabel(['Ratio']);

subplot(3,1,2);
scatter(ti,x,10,'+'); hold on;
plot(3*(1:k),y(:,ir(2)),'+-');
plot([-620,-620], [-1.5,1.5], 'r');
plot([-417,-417], [-1.5,1.5], 'r');
plot([-336,-336], [-1.5,1.5], 'r');
plot([-129,-129], [-1.5,1.5], 'r');
hold off; eval(xa); xlim(3*[-nx k+1]); ylim([-1.5,1.5]);
ylabel(['Ratio']);

subplot(3,1,3);
scatter(ti,x,10,'+'); hold on;
plot(3*(1:k),y(:,ir(3)),'+-');
plot([-620,-620], [-1.5,1.5], 'r');
plot([-417,-417], [-1.5,1.5], 'r');
plot([-336,-336], [-1.5,1.5], 'r');
plot([-129,-129], [-1.5,1.5], 'r');
hold off; eval(xa); xlim(3*[-nx k+1]); ylim([-1.5,1.5]);
ylabel(['Ratio']);
```





When we investigate original data plot at figure3, we can find a specific periodic feature in O2 level. For instance, about 620k, 415k, 330k, and 129kyears ago, we can find sudden and dramatic drop and before these drop, there were steady increase in O2 level with some fluctuations. However, in our synthetic futurea at figure4, there isn't any dramatic drop as much as in our real data even though there are some fluctuations. Thus we can conclude that this is not good enough model and we need other improved model to capture this feature.

1.(c)

i. The time (from now=T) to the next 240,000year maximum level in O2,

There are several options for determining the time that has maximum level in O2. We can choose time that has the largest O2 level in every simulation or time that has largest average O2 level as maximum level time. Among them I choose median and mean value of O2 level as criterion to consider effect of outliers.

```
%Calculate mean and median of synthetic futures
mdn = median(y,2);
mn = mean(y,2);
[max_mdn, where] = max(mdn)
[max_mn, where] = max(mn)


max_mdn =

    0.0808


where =

    26


max_mn =

    0.0835


where =

    25
```

We can check that mean and median has similiar time point that has largest O2 level at 25~26. Thus, we can predict that the time that has maximum level O2 in the next 240,000year is about 75,000 ~ 78,000 years.

ii. As we have seen above, at 25~26, the median O2 level is 0.0808 and mean O2 level is 0.0835. That is I predict O2 level has maximum value which is 0.0835 on average after 75,000 ~ 78,000 years

iii. For the time that has smallest O2 level is 9,000 year(k=3) and it has -0.1342 for median of O2 level and -0.1389 for mean of O2 level. I predict O2 level has minimum level which is -0.1398 on average after 9,000 years

```
[min_mdn, where] = min(mdn)
[min_mn, where] = min(mn)


min_mdn =

   -0.1342
```

```
where =

     3


min_mn =

   -0.1389


where =

     3
```

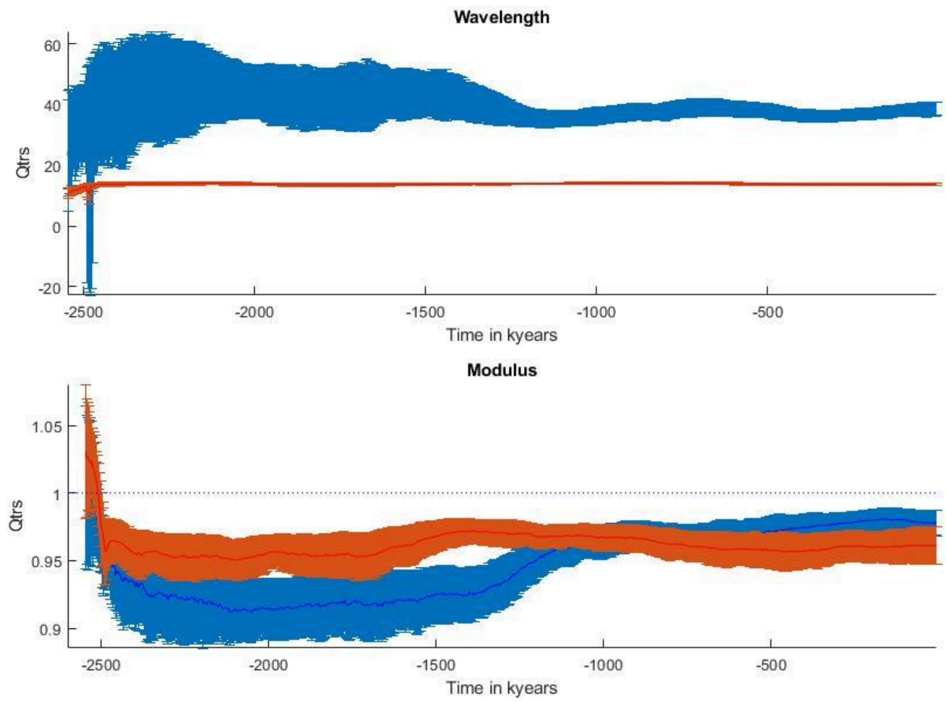
Question 2 - set up

```
rng(642)
I = 1000;  % MC sample size
[thetasamp,vsamp] = tvarFFBS(x,p,del,m0,C0,s0,n0,I);
```

2.(a)

```
nc=2;   % select only the 2 largest wavelength components here
waves=zeros(nc,T,I); mods=waves;
for i=1:I
    [wa,mo] = tvar_decomp(x,thetasamp(:,:,i));
    waves(:,:,i)=wa(1:nc,:); mods(:,:,i)=mo(1:nc,:);
end

    %wavelength plot
figure(5); clf;
pr = prctile(squeeze(waves(1,:,:))',[5 25 50 75 95])';
line(ti(p+1:T),pr(p+1:T,3),'color','b'); hold on;
errorbar(ti(p+1:T),pr(p+1:T,3), pr(p+1:T,4) - pr(p+1:T,3));
pr = prctile(squeeze(waves(2,:,:))',[5 25 50 75 95])';
errorbar(ti(p+1:T),pr(p+1:T,3), pr(p+1:T,4) - pr(p+1:T,3));
line(ti(p+1:T),pr(p+1:T,3),'color','r')
eval(xa);
ylabel('Qtrs'); title(['Wavelength']);

    %moduli plot
figure(6); clf
pr = prctile(squeeze(mods(1,:,:))',[5 25 50 75 95])';
errorbar(ti(p+1:T),pr(p+1:T,3), pr(p+1:T,4) - pr(p+1:T,3));
line(ti(p+1:T),pr(p+1:T,3),'color','b'); hold on;
pr = prctile(squeeze(mods(2,:,:))',[5 25 50 75 95])';
errorbar(ti(p+1:T),pr(p+1:T,3), pr(p+1:T,4) - pr(p+1:T,3));
line(ti(p+1:T),pr(p+1:T,3),'color','r')
axis tight; ylim([0 1.1]); box off; eval(xa)
hold on; plot([ti(1) ti(T)],[1 1],'b:');hold off
```

```
ylabel('Qtrs'); title(['Modulus'])
```



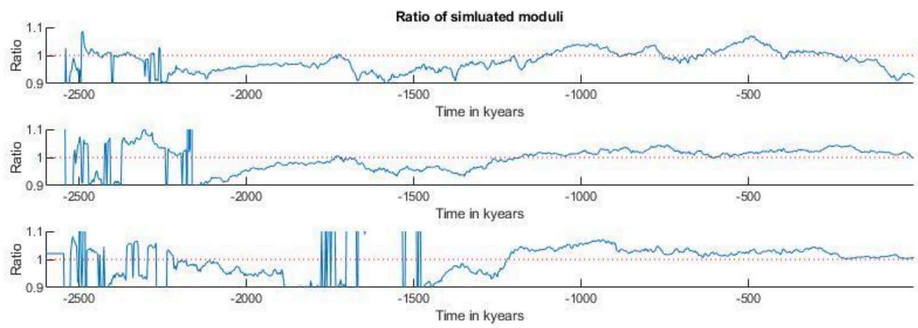
The first component which is indicated by blue line has much large wavelength than second component and this relationship keeps constant over time as we can see in figure 5. However, in moduli plot, we can find that the relationship between 2 components has changed from about 1000 years ago. At first, moduli of second component(smaller wavelength) which is indicated with red line is larger than blue one. But after some time point, blue line becomes larger than red line. This is very interesting in that the dominant component in moduli underlying this process has changed as time went. Consequently, to capture these time varying feature TVAR DLM should be used in this study. Furthermore, in the moduli plot, second component indicated with red line was not locally stationary at about 2500k years ago because moduli had larger value than 1.

2.(b)

i.

```
rng(642);
n=3; in=randsample(I,n);
r = mods(1,:,in)./mods(2,:,in);
figure(7); clf;
subplot(3,1,1);
plot(ti(1:T),r(:,:,1)); hold on;
plot([ti(1) ti(T)],[1 1],'r:'); hold off;
eval(xa); ylabel('Ratio');
ylim([0.9 1.1]);
title(['Ratio of simluated moduli']);  ylabel(['Ratio'])

subplot(3,1,2);
plot(ti(1:T),r(:,:,2)); hold on;
plot([ti(1) ti(T)],[1 1],'r:'); hold off;
eval(xa); ylabel('Ratio');
```

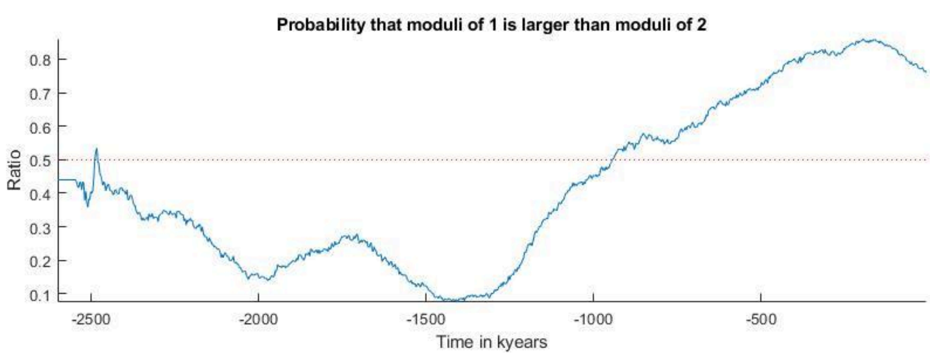
```
ylim([0.9 1.1]);

subplot(3,1,3);
plot(ti(1:T),r(:,:,3)); hold on;
plot([ti(1) ti(T)],[1 1],'r:'); hold off;
eval(xa); ylabel('Ratio');
ylim([0.9 1.1]);
```



We can consider that first component as 110k year periodic component and second component as 40k year periodic component. When we investigate some simulated trajectories of ratio of two moduli, there are some fuctuation in ratio of two moduli. However, from about 1200kyears ago, ratio constantly increased and keeps slightly larger than 1(which is not at first plot from 500kyears ago).

ii.

```
pr = mods(1,:,:)./mods(2,:,:);
pr_sim = mean(pr>1,3);
figure(8); clf;
plot(ti(1:T),pr_sim); hold on;
plot([ti(1) ti(T)],[0.5 0.5],'r:'); hold off;
eval(xa); ylabel('Ratio');
title('Probability that moduli of 1 is larger than moduli of 2')
```



As we can see in the plot, the probability that first component's moduli is larger than second's constantly decreased from beginning to about 1300 years ago which ended up with 0.1. However, after this decrease, there is significant increase and it becomes larger than 0.5 from 1000kyears ago. Finally the probability that moduli of first component is larger than second's is around 77%.

Based on summaries in this question, every plot and summaries are indicating the change of dominant underlying component in this process. Thus we can statistically conclude that the significance of the 110kyears component has gradually increased which result in a significant structure climatic change.

Question 3 - set up

$x_t|\phi_t \sim pois(\phi_t)$

$\phi_{t-1}|D_{t-1} \sim gamma(a_{t-1}, a_{t-1}/m_{t-1})$

$\phi_t = \phi_{t-1}\eta_t/\beta$

3.(a)

In HW5's Exercise 4, we have proven that $\phi_0 \sim gamma(a,b),\ \eta \sim Be(\beta a, (1-\beta)a)$, then, $\phi_1 = \phi_0\eta/\beta \sim gamma(\beta a, \beta b)$

in this case, $\phi_{t-1}|D_{t-1} \sim gamma(a_{t-1}, \frac{a_{t-1}}{m_{t-1}}), \eta_t|D_{t-1} \sim Be(\beta a_{t-1}, (1-\beta)a_{t-1})$. Thus, $P(\phi_t|D_{t-1}) \sim N(\beta a_{t-1}, \frac{\beta a_{t-1}}{m_{t-1}})$

3.(b)

$E(x_t \mid D_{t-1}) = E(E(x_t|\phi_t, D_{t-1})) = E(\phi_t|D_{t-1}) = \frac{\beta a_{t-1}}{\frac{\beta a_{t-1}}{m_{t-1}}} = m_{t-1}$

3.(c)

$$P(x_t|D_{t-1}) = \int P(x_t, \phi_t|D_{t-1})d\phi_t$$

$$= \int P(x_t|\phi_t, D_{t-1})P(\phi_t|D_{t-1})d\phi_t$$

$$= \int P(x_t|\phi_t)P(\phi_t|D_{t-1})d\phi_t$$

$$= \frac{1}{\Gamma(\beta a_{t-1})x_t!} \times (\frac{\beta a_{t-1}}{m_{t-1}})^{\beta a_{t-1}} \int e^{-\phi_t}\phi_t^{x_t}e^{-\phi_t\frac{\beta a_{t-1}}{m_{t-1}}}\phi_t^{\beta a_{t-1}-1}d\phi_t$$

$$= \frac{1}{\Gamma(\beta a_{t-1})x_t!} \times (\frac{\beta a_{t-1}}{m_{t-1}})^{\beta a_{t-1}} \int e^{-\phi_t(1+\frac{\beta a_{t-1}}{m_{t-1}})}\phi_t^{x_t+\beta a_{t-1}-1}d\phi_t$$

$$= \frac{1}{\Gamma(\beta a_{t-1})x_t!} \times (\frac{\beta a_{t-1}}{m_{t-1}})^{\beta a_{t-1}} \times \Gamma(x_t+\beta a_{t-1}) \times (1+\frac{\beta a_{t-1}}{m_{t-1}})^{-(x_t+\beta a_{t-1})}$$

$$= \frac{\Gamma(\beta a_{t-1}+x_t)}{\Gamma(x_t+1)\Gamma(\beta a_{t-1})}(\frac{\frac{\beta a_{t-1}}{m_{t-1}}}{1+\frac{\beta a_{t-1}}{m_{t-1}}})^{\beta a_{t-1}}(1+\frac{\beta a_{t-1}}{m_{t-1}})^{-x_t}$$

$$= \frac{\Gamma(\beta a_{t-1}+x_t)}{\Gamma(x_t+1)\Gamma(\beta a_{t-1})}(\frac{\beta a_{t-1}}{m_{t-1}+\beta a_{t-1}})^{\beta a_{t-1}} \times (\frac{m_{t-1}}{m_{t-1}+\beta a_{t-1}})^{x_t} \sim nb(\beta a_{t-1}, \frac{\beta a_{t-1}}{m_{t-1}+\beta a_{t-1}})$$

3.(d)

$$P(\phi_t|D_t) = P(\phi_t|D_{t-1}, x_t)$$

$\propto P(\phi_t, x_t|D_{t-1})$

$\propto P(x_t|\phi_t)P(\phi_t|D_{t-1})$

$\propto \phi_t^{x_t}e^{-\phi_t} \times \phi_t^{\beta a_{t-1}-1}e^{-\phi_t \frac{\beta a_{t-1}}{m_{t-1}}}$

$\propto \underbrace{\phi_t^{\beta a_{t-1}+x_t-1}e^{-\phi_t(1+\frac{\beta a_{t-1}}{m_{t-1}})}}_{kernel\ of\ gamma(\beta a_{t-1}+x_t, 1+\frac{\beta a_{t-1}}{m_{t-1}})}$

$\rightarrow P(\phi_t|D_t) \sim gamma(\beta a_{t-1} + x_t, 1 + \frac{\beta a_{t-1}}{m_{t-1}}) = gamma(a_t, \frac{a_t}{m_t})$

where $a_t = \beta a_{t-1} + x_t$ and $\frac{a_t}{m_t} = 1 + \frac{\beta a_{t-1}}{m_{t-1}} \rightarrow \frac{m_t}{a_t} = \frac{m_{t-1}}{m_{t-1}+\beta a_{t-1}} \rightarrow m_t = (\frac{\beta a_{t-1}+x_t}{m_{t-1}+\beta a_{t-1}})m_{t-1}$

m_t positively depends on x_t which means that as x_t increase, m_t also increase. we can interpret this as correction procedure. By observing new value, we adjust our estimate of ϕ. In this context, a_{t-1} plays a role as how much weight we give on previous estimate(like prior sample number). Thus, if a_{t-1} is large, correction from new observation becomes weak and we give more weight on prior estimate.

3.(e)

As we have shown HW5's Exercise 4.(e), $P(\phi_0|\phi_1)$ is defined by $\phi_0 = \beta\phi_1 + \nu$ where $\nu \sim gamma((1-\beta)a, b)$ with $\phi_1 \perp \nu$.

When we adjust above result, we get $\phi_{t-1} = \beta\phi_t + \nu_t$ where $\nu_t \sim gamma((1-\beta)a_{t-1}, \frac{a_{t-1}}{m_{t-1}}$ which indicates first order markovian structure.

Since ϕ_t has first order markovian structure, $\phi_{t-1} \leftarrow \phi_t \leftarrow \cdots \phi_T$.

We can simulate retrospective distribution for $\phi_{T-1}|\phi_T, D_T$ as follow:

From $\phi_T|D_T \sim gamma(a_T, a_T/m_T)$, smooth retrospective distribution for $t = T_1, \cdots 1$.

At first, $m_T^* = m_T, a_T^* = a_T$.

Then $E(\phi_{T-1}|D_T) = m_{T-1}^* = \beta E(\phi_T|D_T) + (1-\beta)m_{T-1} = \beta m_T^* + (1-\beta)m_{T-1}$.

$a_{T-1}^* = (1-\beta)a_{T-1} + \beta a_T^*$.

Then $P(\phi_{T-1}|\phi_T, D_T) \sim gamma(a_{T-1}^*, \frac{a_{T-1}^*}{m_{T-1}^*})$.

From above result, by similar procedure, we can simulate retrospective distribution $P(\phi_{t-1}|\phi_t, D_T) \sim gamma(a_{t-1}^*, \frac{a_{t-1}^*}{m_{t-1}^*})$ for $t = T-1, \cdots 1$.

3.(f)

We can simulate full posterior distribution of $\phi_1, \cdots \phi_T$ from the fact $\phi_{t-1} = \beta\phi_t + \nu_t$, where $\nu_t \sim gamma(a_{t-1}, \frac{(1-\beta)a_{t-1}}{m_{t-1}})$.

1. sample $\phi_T^{(s)}$ from $\phi_T|D_T \sim gamma(a_T, \frac{a_T}{m_T})$

2. sample $\nu_T^{(s)}$ from $gamma((1-\beta)a_{T-1}, \frac{a_{T-1}}{m_{T-1}})$

3. calculate $\phi_{T-1}^{(s)} = \beta\phi_T^{(s)} + \nu_T^{(s)}$

4. sample $\nu_{T-1}^{(s)}$ from $gamma((1-\beta)a_{T-2}, \frac{a_{T-2}}{m_{T-2}})$

5. calcuate $\phi_{T-2}^{(s)} = \beta\phi_{T-1}^{(s)} + \nu_{T-1}^{(s)}$

iterate above procedure, then we can get $(\phi_1, \cdots \phi_T)^{(s)}$ simulated full joint posterior distribution.

Question 4 - set up

```
y = load('intrusionevents.txt');
x = y(:,2); ti = y(:,1);
T = length(x);
```

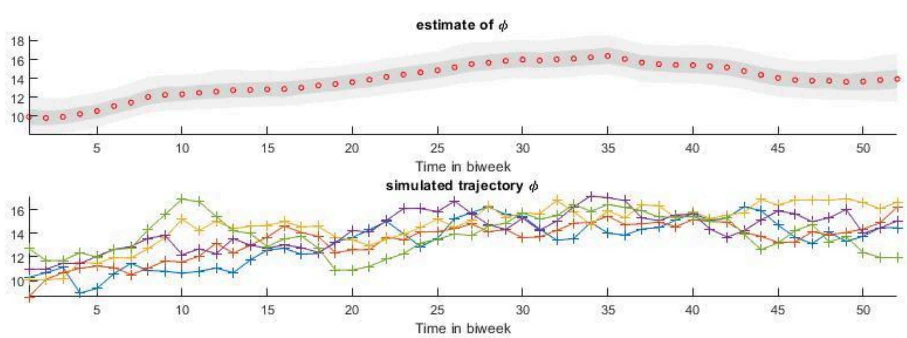
4.(a)

```
%priors
a0 = 25; m0 = 14.5; bn = 0.8:0.01:0.99;
%organize data and allocate space
arx=reshape(x-mean(x),length(x),1);
likp=zeros(1,length(bn));
bopt=1; maxlik=-1e300;
%explore space of b
for j = 1:length(bn)
    b = bn(j);
    mt = m0; at = a0;
    llik = 0;
    for t = 1:T
        %evolve
        at = b*at;
        llik = llik + log(nbinpdf(x(t),at, at/(mt+at)));
        %update
        mt = mt*((at+x(t))/(mt+at)); at = at+x(t);
    end;
    likp(j) = llik;
    if (llik > maxlik)
        bopt = b; maxlik = llik;
    end;
end;
bopt


bopt =

    0.8200
```

We have found that $p(x_t|D_{t-1}) \sim nb(\beta a_{t-1}, \frac{\beta a_{t-1}}{(m_{t-1}+\beta a_{t-1})})$ at 3.(c). From 0.8 to 0.99, I have explored joint log likelihood values which is sum of $log(p(x_t|D_{t-1}))$ for each discount factor values, as a result, I have found that 0.82 is optimal value for discount factor because it has the largest loglikelihood.

4.(b)

```
nmc = 1000;
b = bopt;
a = zeros(1,T); m = zeros(1,T);
at = a0; mt = m0;
for t = 1:T
    at = b*at;
    mt = mt*((at+x(t))/(mt+at)); at = at+x(t);
    m(:,t) = mt; a(:,t) = at;
end;
psamp = zeros(T,nmc);
pt = gamrnd(a(T), m(T)/a(T),1,nmc);
psamp(T,:) = pt;
rng(642)
for t = (T-1):-1:1
    pt = b*pt + gamrnd((1-b)*a(t),m(t)/a(t),1,nmc);
    psamp(t,:) = pt;
end;

figure(9); clf;
subplot(2,1,1);
pr = prctile(squeeze(psamp)',[5 25 50 75 95])';
ciplot(pr(:,1),pr(:,5),ti,[0.95 0.95 0.95]); hold on
ciplot(pr(:,2),pr(:,4),ti,[0.85 0.85 0.85]);
scatter(ti,pr(:,3),10,'ro');
xa='box off; axis tight; xlabel(''Time in biweek''); ylabel('''')';
hold off; eval(xa);
title(['estimate of \phi']);

subplot(2,1,2);
r = 5; ir = randsample(nmc,r);
plot(ti,psamp(:,ir),'+-');
title(['simulated trajectory \phi'])
xa='box off; axis tight; xlabel(''Time in biweek''); ylabel('''')';
eval(xa);
```

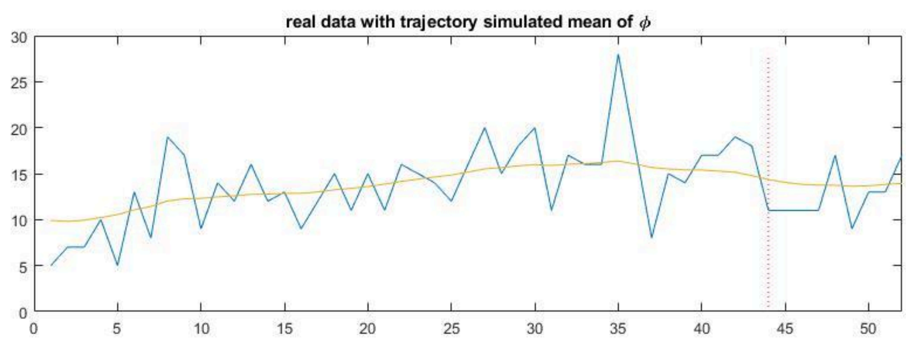


When we investigate plots in figure 9 which shows estimated trajectory ϕ on average , there is an increase from beginning point to 35th point. After this point we can also find gradual decrease to most recent time point. From this fact, we can find existence of time varying mean feature.

4.(c)

```
ratio = psamp(52,:)./psamp(44,:);
mean(ratio>1)
figure(10); clf;
plot(x); hold on;
plot([44 44],[0 max(x)],'r:');
plot(ti,pr(:,3));
xlim([0 52]);
title(['real data with trajectory simulated mean of \phi'])
```

```
ans =

    0.4070
```



In the real data, At 44th time point, we can find sudden drop in the number of intrusion which decrease as much as 7. After this point, data keeps small count than before except for two cases. Thus we can say that modification in system is effective when we see real data. However, simulated ratio of ϕ_{52} and ϕ_{44} does not support this phenomenon. Among 1000 simulated trajectory phi, about 40% says that ϕ_{52} is larger than ϕ_{44}. I think this proportion is too large to support significant decrease of phi. Thus, in my opinion, we cannot provide statistical significance of impact from 44th time point.

4.(d)

Now we have prior $P(\phi_{52}|D_{52}) \sim Gamma(a_t, a_t/m_t)$.

From this prior evolve ϕ and get distribtuion of $P(\phi_{53}|D_{52}) \sim Gamma(\beta a_t, \beta a_t/m_t)$ and sample ϕ_{53}.

Based on ϕ_{53}, forecast x_{53} from $P(x_{53}|\phi_{53}) \sim poisson(\phi_{53})$.

Similarly evolve from $P(\phi_{53}|D_{52})$ to $P(\phi_{54}|D_{52}) \sim Gamma(\beta^2 a_t, \beta^2 a_t/m_t)$ and sample ϕ_{54} and forecast $x_{54} \sim poisson(\phi_{54})$.

In this way we can forecast next K point step by step.

Published with MATLAB® R2019b