10-602: Statistical Approaches to Learning and Discovery

Final Exam Solutions Thomas P Minka

---

1. (a) Let $D_j = \{y_{j1}, ..., y_{jn_j}\}$. The data and parameters are independent for each j. Start by making the standard transformation:

$$\begin{aligned} p(D_j|m_j, v_j) &= \frac{1}{(2\pi v_j)^{n_j/2}} \exp(-\frac{1}{2v_j} \sum_i (y_{ji} - m_j)^2) & (1) \\ &= \frac{1}{(2\pi v_j)^{n_j/2}} \exp(-\frac{1}{2v_j}(S_j + n_j(m_j - \bar{y}_j)^2)) & (2) \\ &= \frac{1}{(2\pi v_j)^{(n_j-1)/2}\sqrt{n_j}} \exp(-\frac{S_j}{2v_j}) \mathcal{N}(m_j; \bar{y}_j, v_j/n_j) & (3) \end{aligned}$$

Now notice that

$$\int_m \mathcal{N}(x; m, v)\mathcal{N}(m; \mu, \alpha)dm = \mathcal{N}(x; \mu, v + \alpha) \quad (4)$$

because $x = m + \mathcal{N}(0, v)$ and $m = \mu + \mathcal{N}(0, \alpha)$. This gives

$$\begin{aligned} p(v_j|D_j) &= p(v_j) \int_{m_j} p(D_j|m_j, v_j)\mathcal{N}(m_j; \mu_j, \alpha_j^{-1}v_j)dm_j & (5) \\ &\propto p(v_j) \frac{1}{v_j^{(n_j-1)/2}} \exp(-\frac{S_j}{2v_j}) \int_{m_j} \mathcal{N}(m_j; \bar{y}_j, v_j/n_j)\mathcal{N}(m_j; \mu_j, \alpha_j^{-1}v_j)dm_j & (6) \\ &\propto p(v_j) \frac{1}{v_j^{(n_j-1)/2}} \exp(-\frac{S_j}{2v_j}) \mathcal{N}(\bar{y}_j; \mu_j, v_j(1/n_j + 1/\alpha_j)) & (7) \end{aligned}$$

If we define

$$\begin{aligned} c_j &= \nu_j + n_j/2 & (8) \\ d_j &= \lambda_j + \frac{S_j}{2} + \frac{(\bar{y}_j - \mu_j)^2}{2(1/n_j + 1/\alpha_j)} & (9) \\ p(v_j|D_j) &\sim IG(c_j, d_j) & (10) \end{aligned}$$

(b)

$$\begin{aligned} E[v_j|D_j] &= \frac{d_j}{c_j - 1} & (11) \\ \mathrm{var}(v_j|D_j) &= \frac{d_j^2}{(c_j - 1)^2(c_j - 2)} & (12) \\ \gamma_j &= \log v_j & (13) \\ \frac{d\gamma_j}{dv_j} &= \frac{1}{v_j} & (14) \\ E[\gamma_j|D_j] &\approx \log E[v_j|D_j] = \log \frac{d_j}{c_j - 1} & (15) \\ \mathrm{var}(\gamma_j|D_j) &\approx \frac{\mathrm{var}(v_j|D_j)}{E[v_j|D_j]^2} = \frac{1}{c_j - 2} & (16) \\ Pr(\gamma_1 < \gamma_2|D) &\approx \phi\left(\frac{\log(d_2/(c_2 - 1)) - \log(d_1/(c_1 - 1))}{\sqrt{1/(c_2 - 2) + 1/(c_1 - 2)}}\right) & (17) \end{aligned}$$

(c)

$$
\begin{aligned}
p(\gamma_j|D_j) &\propto \frac{1}{v_j^{c_j}} \exp(-\frac{d_j}{v_j}) && (18) \\
\frac{d \log p(\gamma_j|D_j)}{d\gamma_j} &= -c_j + d_j e^{-\gamma_j} && (19) \\
\hat{\gamma}_j &= \log \frac{d_j}{c_j} && (20) \\
\frac{d^2 \log p(\gamma_j|D_j)}{d\gamma_j^2} &= -d_j e^{-\gamma_j} && (21) \\
\left.\frac{d^2 \log p(\gamma_j|D_j)}{d\gamma_j^2}\right|_{\gamma_j=\hat{\gamma}_j} &= -c_j && (22) \\
p(\gamma_j|D_j) &\approx \mathcal{N}(\log \frac{d_j}{c_j}, \frac{1}{c_j}) && (23) \\
Pr(\gamma_1 < \gamma_2|D) &\approx \phi\left(\frac{\log(d_2/c_2) - \log(d_1/c_1)}{\sqrt{1/c_2 + 1/c_1}}\right) && (24)
\end{aligned}
$$

(d)

$$
\begin{aligned}
c_1 = 11 \qquad & d_1 = 12955 && (25) \\
c_2 = 7 \qquad & d_2 = 16172 && (26) \\
Pr(\gamma_1 < \gamma_2|D) \approx\ & 0.9055 \text{ (Delta)} && (27) \\
Pr(\gamma_1 < \gamma_2|D) \approx\ & 0.9183 \text{ (Laplace)} && (28)
\end{aligned}
$$

The true value is $Pr(v_1 < v_2|D) = 0.9240$.

2. (a)

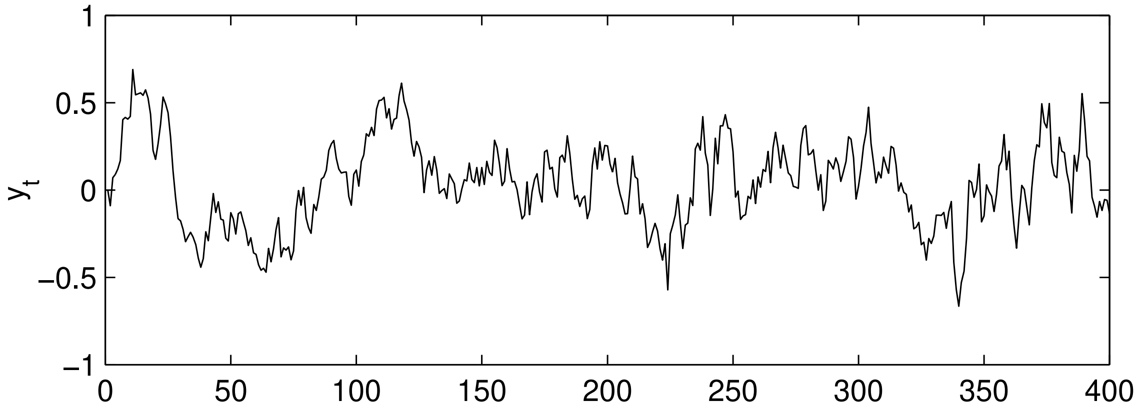
$$
\begin{aligned}
S_{xx} &= \sum_{t=1}^{n} y_{t-1}^2 && (29) \\
S_{yx} &= \sum_{t=1}^{n} y_t y_{t-1} && (30) \\
p(D|a,v) &= \frac{1}{(2\pi v)^{n/2}} \exp(-\frac{1}{2v}\left(a^2 S_{xx} - 2aS_{yx} + \sum_{y=1}^{n} y_t^2\right)) && (31) \\
\hat{a} &= \frac{S_{yx}}{S_{xx}} && (32) \\
\hat{v} &= \frac{1}{n}\sum_{y=1}^{n}(y_t - \hat{a}y_{t-1})^2 && (33)
\end{aligned}
$$

(b)

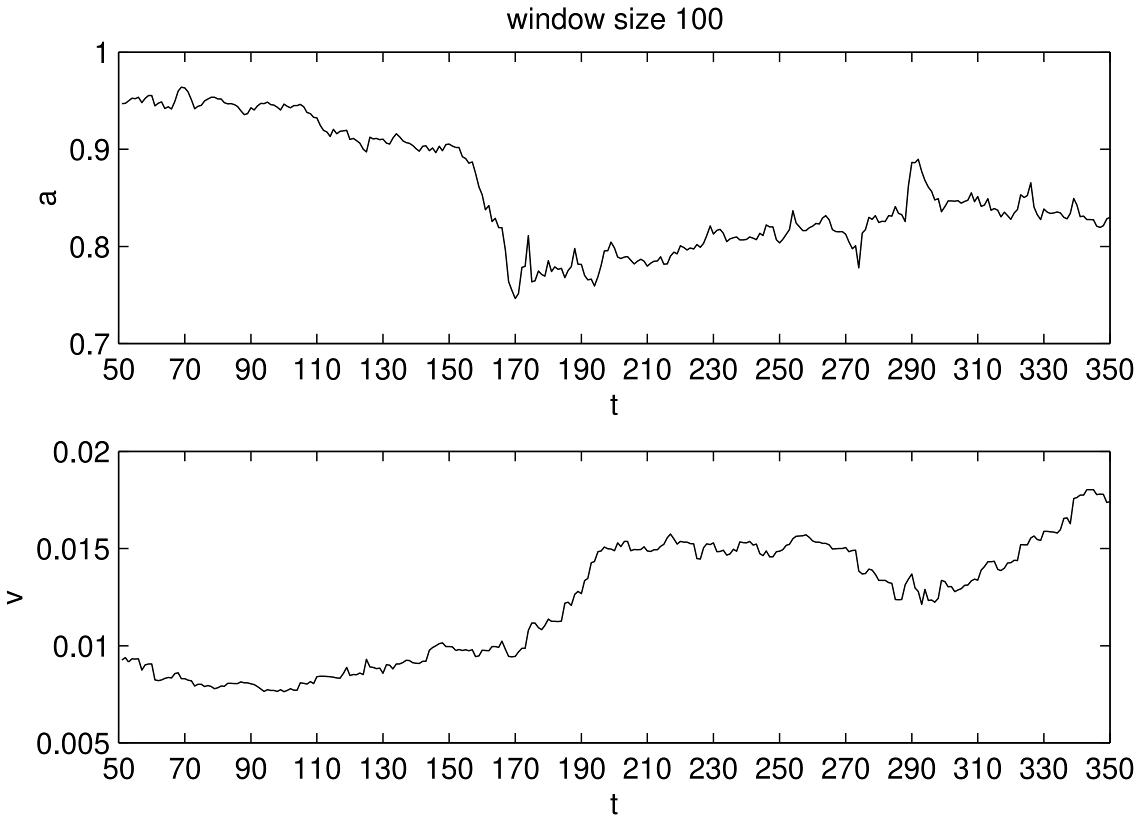
$$
\begin{aligned}
p(a) &\propto \exp(-\frac{1}{2}(a^2/v_a + 2am_a/v_a + m_a^2/v_a)) && (34) \\
p(a|v,D) &\propto \exp(-\frac{1}{2}\left(a^2(S_{xx}/v + 1/v_a) + 2a(S_{yx}/v + m_a/v_a)\right)) && (35)
\end{aligned}
$$

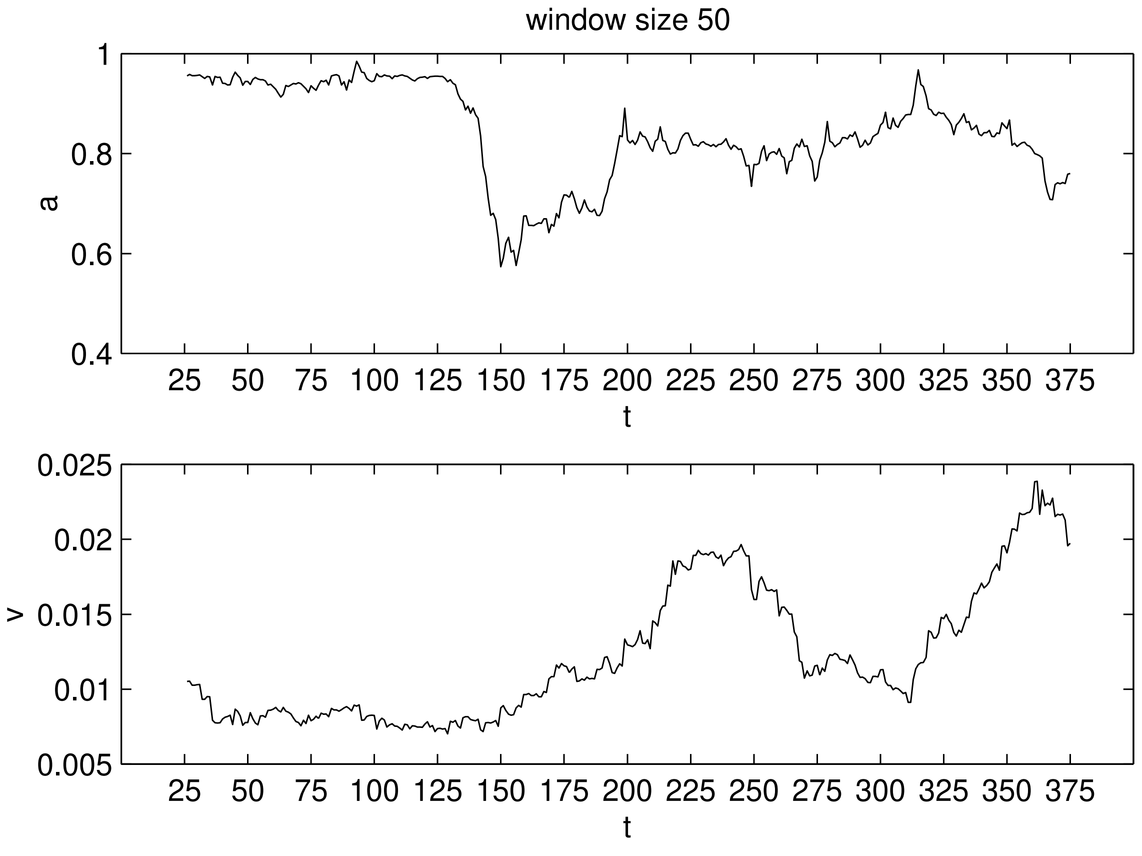
$$
\begin{aligned}
&\sim \quad \mathcal{N}\left(\frac{S_{yx}/v + m_a/v_a}{S_{xx}/v + 1/v_a}, \frac{1}{S_{xx}/v + 1/v_a}\right) & (36)\\
p(v|a, D) &\propto \quad \frac{1}{v^{c+1}v^{n/2}} \exp(-\frac{1}{v}\left(d + \frac{1}{2}\sum_{t=1}^{n}(y_t - ay_{t-1})^2\right)) & (37)\\
&\sim \quad IG(c + n/2, d + \frac{1}{2}\sum_{t=1}^{n}(y_t - ay_{t-1})^2) & (38)
\end{aligned}
$$

3. (a) Plotting the data as a curve does not suggest any single changepoint, though there is an unusual interval from $t = 125$ to $t = 200$:



A better way to see the changepoint is to compute a local MLE $(\hat{a}_t, \hat{v}_t)$ in the neighborhood of each time t. The neighborhood can be defined as $[t - w/2, t + w/2]$ where w is a fixed window size. The plots on the next page show the result for $w = 100$ and $w = 50$. The plot with $w = 100$ strongly suggests a change in a as well as a change in v, though not necessarily at the same time. With $w = 50$, additional changepoints in a and v are suggested. (However, this data is known to be generated with only one changepoint at $t = 166$.)


window size 100
a
1
0.9
0.8
0.7
50 70 90 110 130 150 170 190 210 230 250 270 290 310 330 350
t
v
0.02
0.015
0.01
0.005
50 70 90 110 130 150 170 190 210 230 250 270 290 310 330 350
t


window size 50
a
1
0.8
0.6
0.4
25 50 75 100 125 150 175 200 225 250 275 300 325 350 375
t
v
0.025
0.02
0.015
0.01
0.005
25 50 75 100 125 150 175 200 225 250 275 300 325 350 375
t

(b) The likelihood function is

$$p(D|k, a_1, v_1, a_2, v_2) = \prod_{t=1}^{k} \mathcal{N}(y_t; a_1 y_{t-1}, v_1) \prod_{t=k+1}^{n} \mathcal{N}(y_t; a_2 y_{t-1}, v_2) \quad (39)$$

Given k, the MLEs for a_1, v_1, a_2, and v_2 can be computed using the results of the last problem. So the joint MLE is found by enumerating all k and finding the maximum of $p(D|k, \hat{a}_1(k), \hat{v}_1(k), \hat{a}_2(k), \hat{v}_2(k))$. The result is:

k	166		
a_1	0.945339	a_2	0.809881
v_1	0.00872558	v_2	0.0157403

These values agree with the $w = 100$ plot in part (a).

(c)

$$p(k|D, a_1, v_1, a_2, v_2) = \frac{p(D|k, a_1, v_1, a_2, v_2)p(k)}{\sum_k p(D|k, a_1, v_1, a_2, v_2)p(k)} \quad (40)$$

$$p(k|D) = \int p(k|D, a_1, v_1, a_2, v_2)p(a_1, v_1, a_2, v_2|D)\ da_1\ dv_1\ da_2\ dv_2 \quad (41)$$

So, if we get samples from $p(a_1, v_1, a_2, v_2|D)$ we can approximate

$$p(k|D) \approx \frac{1}{s}\sum_s p(k|D, a_1^{(s)}, v_1^{(s)}, a_2^{(s)}, v_2^{(s)}) \quad (42)$$

These samples can be obtained by Gibbs sampling from the full posterior $p(m_a, v_a, k, a_1, v_1, a_2, v_2|D)$ and discarding the (m_a, v_a, k) values. Another method, which is less efficient, is to get samples from $p(k|D)$ and make a histogram:

$$p(k|D) \approx \frac{1}{s}\sum_s \delta(k = k^{(s)}) \quad (43)$$

(d)

$$p(a_1|D, k, v_1, m_a, v_a) \sim \mathcal{N}\left(\frac{\frac{\sum_{t=1}^{k} y_t y_{t-1}}{v_1} + \frac{m_a}{v_a}}{\frac{\sum_{t=1}^{k} y_{t-1}^2}{v_1} + \frac{1}{v_a}}, \frac{1}{\frac{\sum_{t=1}^{k} y_{t-1}^2}{v_1} + \frac{1}{v_a}}\right) \quad (44)$$

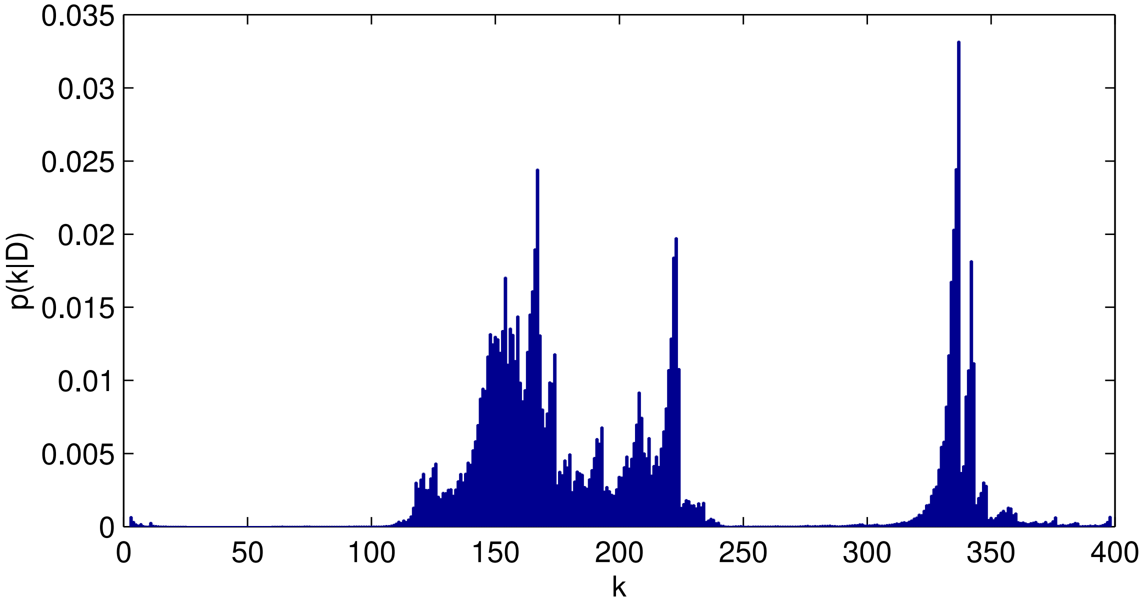
$$p(v_1|D, k, a_1, c, d) \sim IG(c + k/2, d + \frac{\sum_{t=1}^{k}(y_t - a_1 y_{t-1})^2}{2}) \quad (45)$$

$$p(m_a|D, a_1, a_2) \sim \mathcal{N}(\frac{a_1 + a_2}{2}\frac{\frac{2}{v_a}}{\frac{2}{v_a} + 1}, \frac{1}{\frac{2}{v_a} + 1}) \quad (46)$$

$$p(v_a|D, a_1, a_2) \sim IG(2, 1 + \frac{(a_1 - m_a)^2 + (a_2 - m_a)^2}{2}) \quad (47)$$

$p(a_2|D, k, v_2, m_a, v_a)$ and $p(v_2|D, k, a_2, c, d)$ are analogous. $p(k|D, a_1, v_1, a_2, v_2, m_a, v_a)$ is given by (40).

(e) Running the sampler for 10,000 iterations starting from the MLEs gives the following estimate of $p(k|D)$:



The peak at $k = 337$, which was not present in the likelihood plot, is due to the $IG(2, 0.1)$ prior on v_1 and v_2. For example, by changing d you can make either peak higher than the other. This illustrates the danger in using "convenient" priors which are too dogmatic.