UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Exam in:	STK4051 — Computational statistics
Day of examination:	Thursday November 30 2017.
Examination hours:	09.00 – 13.00.
This problem set consists of 5 pages.	
Appendices:	None
Permitted aids:	None

Please make sure that your copy of the problem set is complete before you attempt to answer anything.

Problem 1

(a)

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n}[\sum_{k=1}^{K} \pi_k \frac{\lambda_k^{y_i}}{y_i!} e^{-\lambda_k}]$$

$$l(\boldsymbol{\theta}) = \text{Const} + \sum_{i=1}^{n} \log[\sum_{k=1}^{K} \pi_k \lambda_k^{y_i} e^{-\lambda_k}]$$

(b) Nelder-Mead: With $\boldsymbol{\theta}$ p-dimensional, we start with $p+1$ values of $\boldsymbol{\theta}$. These $p+1$ values are dynamically altered by changing the worst value with a better one, defined through a search line going through the worst value and defined by the average of the other values. This method does not need the derivatives

The Quasi-Newton method is a Newton-type method which avoids the calculation of the second derivative (the Hessian) by replacing it with another quantity M that is sequentially updated.

Neither of thee methods are guaranteed to converge to the global optimum, so running it with different starting values is always a good idea. Clearly both methods have problems with local optima in this case.

The Nelder-Mead method seems to be more stable in this case, but on the other hand, the Quasi-Newton method gives to overall best value (although not very different). The Nelder-Mead method is probably preferable in this case.

(Continued on page 2.)

(c) The main idea behind the EM algorithm is that while the marginal likelihood based on the observed values only may be difficult to optimise, the complete likelihood including some hidden variables may be easier to handle. The E-step corresponds to estimating the complete log-likelihood by

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = E[l(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{C})|\boldsymbol{y}\boldsymbol{\theta}^{(t)}]$$

where $l(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{C})$ is the complete log-likelihood based on both $\boldsymbol{y}$ and $\boldsymbol{C}$. The M-step corresponds to optimising $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ with respect to $\boldsymbol{\theta}$.

For this problem, the C_i's are the missing variables, and if these were known the problem would be considerably easier. Therefore the EM-algorithm is well-suited in this case.

(d) The complete log-likelihood corresponds to defining $P(C_i, Y_i)$ for each observation, giving

$$\begin{aligned}
L(\boldsymbol{\theta}) =& \prod_{i=1}^{n} \pi_{c_i} \frac{\lambda_{c_i}^{y_i} \exp(-\lambda_{c_i})}{y_i!} \\
l(\boldsymbol{\theta}) =& \sum_{i=1}^{n} [\log \pi_{c_i} + y_i \log(\lambda_{c_i}) - \lambda_{c_i} - \log(y_i!)] \\
=& \text{Const} + \sum_{i=1}^{n} \sum_{k=1}^{K} I(c_i = k)[\log(\pi_k) + y_i \log(\lambda_k) - \lambda_k] \\
Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) =& E[\text{Const} + \sum_{i=1}^{n} \sum_{k=1}^{K} I(C_i = k)[\log(\pi_k) + y_i \log(\lambda_k) - \lambda_k|\boldsymbol{y}, \boldsymbol{\theta}^{(t)}] \\
=& \text{Const} + \sum_{i=1}^{n} \sum_{k=1}^{K} \Pr(C_i = k|\boldsymbol{y}, \boldsymbol{\theta}^{(t)})[\log(\pi_k) + y_i \log(\lambda_k) - \lambda_k]
\end{aligned}$$

(e) Optimisation of $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ can be obtained by looking at the derivatives. Note however that we have a constraint $\sum_k \pi_k = 1$, so we need to introduce a lagrange term:

$$\frac{\partial}{\partial \pi_k} Q_{larg}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \sum_{i=1}^{n} \Pr(C_i = k|\boldsymbol{y}, \boldsymbol{\theta}^{(t)}) \frac{1}{\pi_k} - \phi$$

giving

$$\pi_k^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} \Pr(C_i = k|\boldsymbol{y}, \boldsymbol{\theta}^{(t)})$$

(Continued on page 3.)

after proper normalisation. Similarly,

$$\frac{\partial}{\partial \lambda_k} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \sum_{i=1}^{n} \Pr(C_i = k|\boldsymbol{y}, \boldsymbol{\theta}^{(t)})[\tfrac{y_i}{\lambda_k} - 1]$$

giving

$$\lambda_k^{(t+1)} = \frac{\sum_{i=1}^{n} \Pr(C_i = k|\boldsymbol{y}, \boldsymbol{\theta}^{(t)}) y_i}{\sum_{i=1}^{n} \Pr(C_i = k|\boldsymbol{y}, \boldsymbol{\theta}^{(t)})}$$

(f) The EM-algorithm is guaranteed to increase the (log)-likelihood value at each iteration, which we see happens from the plot.

In order to obtain uncertainty measures, different possibilities are available:

- Bootstrapping
- Deriving the Hessian, perhaps by one of the direct optimisation routines using that Hessians can directly be evaluated (but then starting the optimisation at the optimum obtained).
- Use some of the methods attached to the EM algorithm for deriving the variance.

Problem 2

(a)

$$\begin{aligned}\int_{\boldsymbol{x}} \pi(\boldsymbol{x}) P(\boldsymbol{y}|\boldsymbol{x}) d\boldsymbol{x} &= \int_{\boldsymbol{x}} \pi(\boldsymbol{y}) P(\boldsymbol{x}|\boldsymbol{y}) d\boldsymbol{x} \\ &= \pi(\boldsymbol{y}) \int_{\boldsymbol{x}} P(\boldsymbol{x}|\boldsymbol{y}) d\boldsymbol{x} = \pi(\boldsymbol{y})\end{aligned}$$

(b) We have

$$\begin{aligned}\pi(\boldsymbol{x}) P(\boldsymbol{y}|\boldsymbol{x}) &= \pi(\boldsymbol{x}) g(\boldsymbol{y}|x) \min\left\{1, \frac{\pi(\boldsymbol{y}) g(\boldsymbol{x}|\boldsymbol{y})}{f(\boldsymbol{x}) g(\boldsymbol{y}|\boldsymbol{x})}\right\} \\ &= \min\{\pi(\boldsymbol{x}) g(\boldsymbol{y}|\boldsymbol{x}), \pi(\boldsymbol{y}) g(\boldsymbol{x}|\boldsymbol{y})\} \\ &= \pi(\boldsymbol{y}) g(\boldsymbol{x}|\boldsymbol{y}) \min\left\{\frac{\pi(\boldsymbol{x}) g(\boldsymbol{y}|\boldsymbol{x})}{\pi(\boldsymbol{y}) g(\boldsymbol{x}|\boldsymbol{y})}, 1\right\} = \pi(\boldsymbol{y}) P(\boldsymbol{x}|\boldsymbol{y})\end{aligned}$$

Also need irreducibility, that is it is possible to move from any state $\boldsymbol{x}$ to any other state $\boldsymbol{y}$ in a finite number of steps. You also need the chain to be aperiodic, but this will be fulfilled as long as there is a positive probability for not accepting a new proposal (which will always be the case except for some degerate situations)

(Continued on page 4.)

(c) For g_1:

$$R = \frac{\sin^2(x^*)\sin^2(2x^*)\phi(x^*;0,1)\phi(x;0,\sigma_1)}{\sin^2(x)\sin^2(2x)\phi(x;0,1)\phi(x^*;0,\sigma_1)}$$

For g_2 we obtain

$$R = \frac{\sin^2(x^*)\sin^2(2x^*)\phi(x^*;0,1)\phi(x;x^*,\sigma_1)}{\sin^2(x)\sin^2(2x)\phi(x;0,1)\phi(x^*;x,\sigma_1)} = \frac{\sin^2(x^*)\sin^2(2x^*)\phi(x^*;0,1)}{\sin^2(x)\sin^2(2x)\phi(x;0,1)}$$

due to the symmetry in the proposal distribution.

The g_1 proposal corresponds to an independent sampler while g_2 corresponds to random walk. For the first one we would like the acceptance rate to be as large as possible, indicating that it is too small in this case. One can try to change σ_1 to see when the acceptance probability is largest. Given that the standard gaussian distribution is involved in the target distribution, something closer to this distribution should be expected to give higher acceptance rate.

For g_2, we want the acceptance rate to be somewhere between 0.25 and 0.50, which is ok in this case (perhaps we could increase the acceptance rate somewhat by decreasing the variance).

(d) A Markov chain typically needs some time before the simulated values are close to the target distribution. In order to reduce the bias, the first iterations should therefore be discarded.

Two possible methods for specifying the burnin:

- Looking at the trace plots and see if they have stabilized
- Calculate the Gelman-Rubin criterion, which, when running multiple chains, mainly compare within variability with between variability. This is a more formal criterion.

(e) If one wants to estimate $E[x^2]$, then the variance of the Monte Carlo estimate converges towards $\sigma^2[1 + \sum_{k=1}^{\infty}\rho(k)]$ where σ^2 corresponds to the variance for independent samples. The two proposal distributions give almost similar estimates on the second part, with a small preference to g_1.

One can also look at the effective sample size which is defined as

$$\frac{L}{1 + \sum_{k=1}^{\infty}\rho(k)}$$

where L is the number of samples used.

(Continued on page 5.)

(f) We have that

$$\begin{aligned}\frac{\pi(x)}{g_1(x)} =& c\frac{\sin^2(x)\sin^2(2x)\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}}{\frac{1}{\sqrt{2\pi}\sigma_1}e^{-\frac{1}{2\sigma_1^2}x^2}} \\ =& c\sigma_1\sin^2(x)\sin^2(2x)e^{-\frac{1}{2}[1-\frac{1}{\sigma_1^2}]x^2} \leq c\sigma_1\end{aligned}$$

showing that the ratio has a finite maximum. The requirements needed are then fulfilled.

(g) For the M-H with g_1, the effective number of samples is estimated to be $10\,000/4.25 \approx 2930$, approximately similar to the number of samples generated by the rejection method. This indicates that the variance will be quite similar.

However, while M-H needed $10000 + 1000 = 11000$ samples to be generated, the rejection sampling required $3000 * 16.9 = 50700$ samples, indicating that the computational effort with rejection sampling was much larger.

An argument towards rejection sampling compared to MCMC is however that the former is exact!