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.

Some distributions that will be used:

$$\text{Poisson}(z;\lambda) = \frac{z^{\lambda}e^{-\lambda z}}{z!} \qquad \text{Poisson}$$

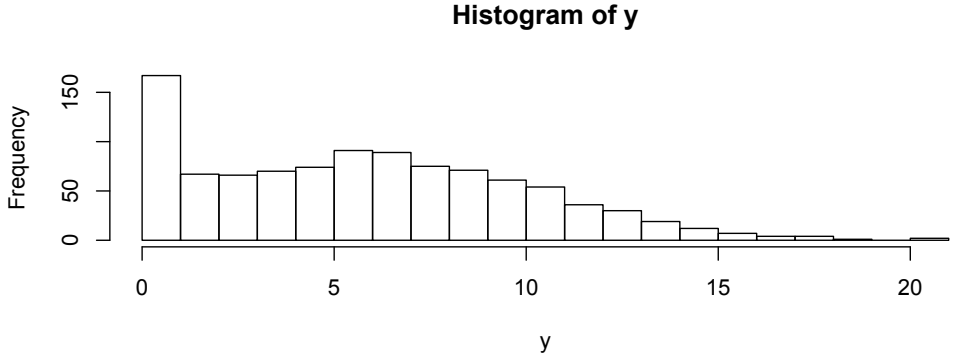
$$\phi(z;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \qquad \text{Gaussian}$$

Problem 1

Assume a mixture model given by

$$Y|C = k \sim \text{Poisson}(y;\lambda_k)$$
$$\Pr[C = k] = \pi_k$$

where $k = 1, ..., K$. We will consider methods for finding the maximum likelihood estimates for $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\lambda})$ based on independent observations $\boldsymbol{y} = (y_1, ..., y_n)$ where $\boldsymbol{\pi} = (\pi_1, ..., \pi_K)$ and $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_K)$.
The plot below shows a histogram of $n = 1000$ observations of Y simulated from the distribution above with $K = 4$.



(Continued on page 2.)

(a) Derive the likelihood function for $\boldsymbol{\theta}$ and show that the log-likelihood function is equal to

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

where Const is a constant not depending on $\boldsymbol{\theta}$

(b) Direct optimization based on the Nelder-Mead algorithm and a quasi-Newton algorithm with 10 random initial values of $\boldsymbol{\theta}$ gave the following best values of l:

Nelder-Mead	-2753.68	-2764.63	-2763.10	-2788.71	-2768.36
	-2755.98	-2753.41	-2765.32	-2758.40	-2758.09
Quasi-Newton	-2753.27	-3274.28	-3274.28	-3274.28	-2753.24
	-2753.40	-2753.24	-3274.28	-2753.55	-2753.24

Describe *briefly* the main concepts behind these methods.

Discuss why it is reasonable to run the optimization algorithms with different starting points.

Based on the results above, which method would you prefer?

(c) An alternative to direct optimization is the EM-algorithm. Describe *briefly* the main idea behind the EM-algorithm and in particular what the $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ function is.

Why is the method suited for this particular problem?

(d) Derive the *complete* likelihood in this case and show that

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \Pr(C_i = k|\boldsymbol{y}, \boldsymbol{\theta}^{(t)})[\log(\pi_k) + y_i \log(\lambda_k) - \lambda_k]$$

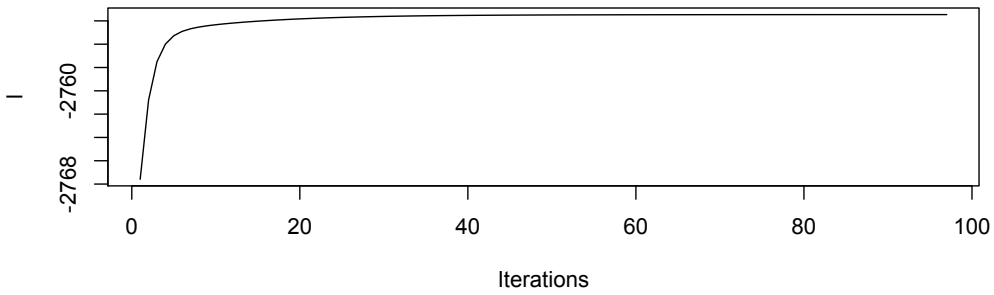
(e) Show that the EM algorithm corresponds to the following updating equations for $\boldsymbol{\pi}$:

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

where you should properly describe $\Pr(C_i = k|\boldsymbol{y}, \boldsymbol{\theta}^{(t)})$. Derive similar equations for $\boldsymbol{\lambda}$.

(f) Running the EM algorithm with the same 10 random initial values of $\boldsymbol{\theta}$ as above all gave the same optimal value of l, equal to -2753.46. Below is a plot of the values of $l(\boldsymbol{\theta}^{(t)})$ as a function of t for one of these runs.

(Continued on page 3.)



Discuss this behaviour related to the general properties of the EM-algorithm.

The following estimates were obtained:

k	1	2	3	4
$\hat{\pi}_k$	0.20	0.17	0.30	0.32
$\hat{\lambda}_k$	0.99	4.01	6.98	10.05

Describe *briefly* possible approaches for obtaining uncertainty measures for the estimates.

Problem 2

Consider the Metropolis-Hastings algorithm where the transition densities $P(\boldsymbol{y}|\boldsymbol{x})$ are defined through:

- Sample a candidate value $\boldsymbol{X}^*$ from a *proposal distribution* $g(\cdot|\boldsymbol{x})$.
- Compute the Metropolis-Hastings ratio

$$R(\boldsymbol{x}, \boldsymbol{X}^*) = \frac{f(\boldsymbol{X}^*)g(\boldsymbol{x}|\boldsymbol{X}^*)}{f(\boldsymbol{x})g(\boldsymbol{X}^*|\boldsymbol{x})}$$

- Put

$$\boldsymbol{Y} = \begin{cases} \boldsymbol{X}^* & \text{with probability } \min\{1, R(\boldsymbol{x}, \boldsymbol{X}^*)\}; \\ \boldsymbol{x} & \text{otherwise.} \end{cases}$$

An essential requirement for a Markov chain to converge to a stationary distribution $\pi(\boldsymbol{x})$ is that

$$\pi(\boldsymbol{y}) = \int_{\boldsymbol{x}} \pi(\boldsymbol{x}) P(\boldsymbol{y}|\boldsymbol{x}) d\boldsymbol{x}. \qquad (*)$$

(Continued on page 4.)

(a) Show that a *sufficient* requirement for (*) is the *detailed balance* criterion

$$\pi(\boldsymbol{y})P(\boldsymbol{x}|\boldsymbol{y}) = \pi(\boldsymbol{x})P(\boldsymbol{y}|\boldsymbol{x}). \tag{1}$$

(b) Show that the Metropolis-Hastings algorithm satisfies the detailed balance criterion.

What other criteria are needed in order for the Markov chain to converge in distribution to $\pi(\boldsymbol{x})$?

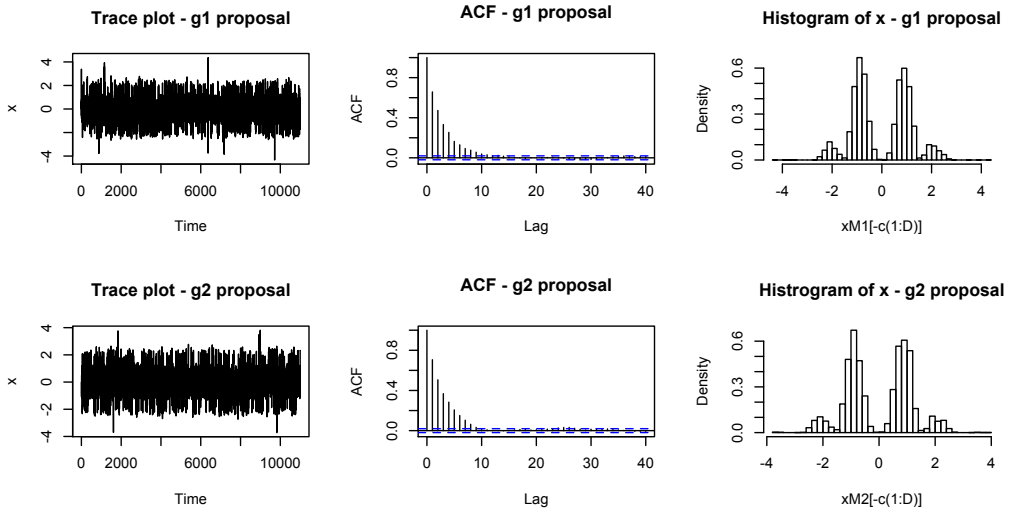
Assume now you want to use the Metropolis-Hastings algorithm to simulate from a one-dimensional distribution given by

$$\pi(x) \propto \sin^2(x) \cdot \sin^2(2x) \cdot \phi(x)$$

where $\phi(x)$ is the density for the standard Gaussian distribution. We will consider two different proposal distributions:

$$g_1(x^*|x) = N(0, \sigma_1^2);$$
$$g_2(x^*|x) = N(x, \sigma_2^2).$$

These two proposal distributions were run with $\sigma_1 = 3.5$ and $\sigma_2 = 2.5$. A total of 11000 iterations were run where the first 1000 were discarded. The plots below shows traceplots (including the first 1000 iterations), estimated autocorrelation functions and histograms for the two proposal distributions. The acceptance rates for the two proposal distributions were 0.24 and 0.28, respectively.



(c) For each of the two proposal distributions, derive formulas for the Metropolis-Hastings acceptance probabilities.

What type of Metropolis-Hastings algorithms do the g_1 and g_2 proposal distributions belong to?

Are the acceptance rates satisfactory for the two proposal distributions? If not, how would you recommend to change the proposal distributions?

(Continued on page 5.)

(d) Why would one discard the values obtained from the first iterations? Describe a general method for specifying the number of iterations to discard.

(e) For the two proposal distributions, the quantity $\sum_{k=0}^{\infty} \rho(k)$, with $\rho(k) = \text{cor}[(x^t)^2, (x^{t+k})^2]$ was estimated to be 3.36 and 3.73, respectively, for the g_1 and g_2 proposal distributions. How can these numbers be used to compare the two versions of the Metropolis-Hasting algorithm if estimation of $E[x^2]$ is of interest?

An alternative to the use of the Metropolis-Hastings algorithm in this case is the rejection sampling method. Assume $g_1(x)$ again is used as proposal distribution.

(f) Does this proposal distribution meet the requirements needed for constructing a proper rejection sampling algorithm?

(g) $N = 3000$ values of x were generated by this algorithm, giving a mean number of proposals equal to 16.9.

In order to estimate $E[x^2]$, would you prefer to use one of the Metropolis-Hastings algorithms or the rejection sampling algorithm. Justify your answer.