STK4051/9051 - Computational statistics - solutions

Trial exam spring 2019

Exercise 1 (a) The inversion method is to generate $X = F_0^{-1}(U)$ where $U \sim \text{Uniform}[0,1]$. We have

$$\begin{aligned}1-e^{-x^\alpha}=&u\\ &\Updownarrow\\ x^\alpha=&-\log(1-u)\\ &\Updownarrow\\ x=&[-\log(1-u)]^{1/\alpha}\end{aligned}$$

(b) We have $x_0 = x/\beta$ giving that

$$\begin{aligned}f(x)=&f_0(x/\beta)/\beta\\ =&\alpha(x/\beta)^{\alpha-1}e^{-(x/\beta)^\alpha}/\beta\\ =&\frac{\alpha x^{\alpha-1}}{\beta^\alpha}e^{-(x/\beta)^\alpha}\end{aligned}$$

showing the result. We can then generate x by

$$x = \beta[-\log(1-u)]^{1/\alpha}$$

(c) We have

$$\begin{aligned}\Pr(X\le x)=&\Pr(F_0^{-1}(\Phi(Y))\le x)\\ =&\Pr(\Phi(Y)\le F_0(x))\\ =&\Pr(Y\le\Phi^{-1}(F_0(x))\\ =&\Phi(\Phi^{-1}(F_0(x)))=F_0(x)\end{aligned}$$

(d) We can then put

$$x_j = F_0^{-1}(\Phi(y_j)).$$

Since (y_1, y_2) are dependent, so will (x_1, x_2) be.

Exercise 2 (a) We will show that the Markov chain satisfies the detailed balance criterion. We have for $\boldsymbol{x} \neq \boldsymbol{y}$

$$\begin{aligned}\pi(\boldsymbol{x})P(\boldsymbol{x},\boldsymbol{y}) =&\pi(\boldsymbol{x})K(\boldsymbol{x},\boldsymbol{y})r_B(\boldsymbol{x},\boldsymbol{y})\\ =&\pi(\boldsymbol{x})K(\boldsymbol{x},\boldsymbol{y})\frac{\pi(\boldsymbol{y})K(\boldsymbol{y},\boldsymbol{x})}{\pi(\boldsymbol{x})K(\boldsymbol{x},\boldsymbol{y})+\pi(\boldsymbol{y})K(\boldsymbol{y},\boldsymbol{x})}\\ =&\pi(\boldsymbol{y})K(\boldsymbol{y},\boldsymbol{x})\frac{\pi(\boldsymbol{x})K(\boldsymbol{x},\boldsymbol{y})}{\pi(\boldsymbol{x})K(\boldsymbol{x},\boldsymbol{y})+\pi(\boldsymbol{y})K(\boldsymbol{y},\boldsymbol{x})}\\ =&\pi(\boldsymbol{y})P(\boldsymbol{y},|\boldsymbol{x}).\end{aligned}$$

(b) If we simulate $\{\boldsymbol{x}^t\}$ according to the described Markov chain, we have from general theory that we can estimate $\mu = E_\pi[h(\boldsymbol{x})]$ by

$$\hat{\mu} = \frac{1}{L}\sum_{t=D+1}^{D+L} h(\boldsymbol{x}^t)$$

where we discard the first D samples in order to minimized the bias due to that it can take some time until the samples are close enough to the target distribution. We further have

$$\begin{aligned}\text{Var}[\hat{\mu}] =&\frac{1}{L^2}[\sum_{t=D+1}^{D+L}\text{Var}[h(\boldsymbol{x}^t)] + 2\sum_{s=D+1}^{D+L-1}\sum_{t=s+1}^{D+L}\text{Cov}[h(\boldsymbol{x}^s),h(\boldsymbol{x}^t)]\\ \approx&\frac{\sigma_h^2}{L}[1+2\sum_{t=D+1}^{D+L-1}\rho(t-s)]\end{aligned}$$

showing the dependence on the correlation structure.

(c) Assume $\pi(\boldsymbol{y})K(\boldsymbol{y},\boldsymbol{x}) \geq \pi(\boldsymbol{x})K(\boldsymbol{x},\boldsymbol{y})$. Then $r_M(\boldsymbol{x},\boldsymbol{y}) = 1 > r_B(\boldsymbol{x},\boldsymbol{y})$. Assume now $\pi(\boldsymbol{y})K(\boldsymbol{y},\boldsymbol{x}) < \pi(\boldsymbol{x})K(\boldsymbol{x},\boldsymbol{y})$. Then

$$r_M(\boldsymbol{x},\boldsymbol{y}) =\frac{\pi(\boldsymbol{y})K(\boldsymbol{y},\boldsymbol{x})}{\pi(\boldsymbol{x})K(\boldsymbol{x},\boldsymbol{y})} \geq \frac{\pi(\boldsymbol{y})K(\boldsymbol{y},\boldsymbol{x})}{\pi(\boldsymbol{x})K(\boldsymbol{x},\boldsymbol{y})+\pi(\boldsymbol{y})K(\boldsymbol{y},\boldsymbol{x})} = r_M(\boldsymbol{x},\boldsymbol{y})$$

showing the first result.

From the general result we then have that $P_M(\boldsymbol{x},\boldsymbol{y}) \geq P_B(\boldsymbol{x},\boldsymbol{y})$ for all $\boldsymbol{x} \neq \boldsymbol{y}$ showing the second result.

Both algorithms are using the same proposals and both have the same invariant distribution. Since M-H give higher acceptance probabilities, the changes should happen more frequent and thereby give a more efficient algorithm.

Exercise 3 (a) We then have

$$\begin{aligned}X|Y \sim&N(\rho Y, 1-\rho^2)\\ Y|X \sim&N(\rho X, 1-\rho^2)\end{aligned}$$

or

$$\begin{aligned}X|Y =&\rho Y + \sqrt{1-\rho^2}Z\\ Y|X =&\rho X + \sqrt{1-\rho^2}V\end{aligned}$$

which, when using the Gibbs sampler gives the recursion

$$\begin{aligned} Y_n =&\rho X_n + \sqrt{1-\rho^2} Z_n, \\ X_n =&\rho Y_{n-1} + \sqrt{1-\rho^2} V_n. \end{aligned}$$

were all the $\{Z_n\}$ and $\{V_n\}$ are independent variables

(b) From the equation above we have

$$\begin{aligned} X_n =&\rho Y_{n-1} + \sqrt{1-\rho^2} V_n \\ =&\rho[\rho X_{n-1} + \sqrt{1-\rho^2} Z_{n-1}] + \sqrt{1-\rho^2} V_n \\ =&\rho^2 X_{n-1} + \sqrt{1-\rho^2}[\rho Z_{n-1} + V_n] \\ =&\rho^2 X_{n-1} + \varepsilon_n \end{aligned}$$

were $E[\varepsilon_n] = 0$ and

$$\text{Var}[\varepsilon_n] =(1-\rho^2)[\rho^2+1] = 1-\rho^4 \equiv \sigma_\varepsilon^2$$

Since $|\rho^2| < 1$, the general results about AR(1) processes applies.

(c) We have

$$E[X_n] =E[\rho^2 X_{n-1} + \varepsilon_n | X_{n-1}] = \rho^2 E[X_{n-1}]$$

which recursively gives $E[X_n] = \rho^{2n}\mu_0$ were $\mu_0 = E[X_0]$.

(d) We have

$$\begin{aligned} \text{Var}[X_n] =&\text{Var}[\rho^2 X_{n-1} + \varepsilon_n] \\ =&\rho^4 \text{Var}[X_{n-1}] + \sigma_\varepsilon^2 \end{aligned}$$

Assume now the statement about the variance is true for n. Then

$$\begin{aligned} \text{Var}[X_{n+1}] =&\rho^4 \text{Var}[X_n] + \sigma_\varepsilon^2 \\ =&\rho^4 \tfrac{\sigma_\varepsilon^2}{1-\rho^4}(1-\rho^{4n}) + \sigma_\varepsilon^2 \\ =&\sigma_\varepsilon^2 \tfrac{\rho^4(1-\rho^{4n})+1-\rho^4}{1-\rho^4} \\ =&\sigma_\varepsilon^2 \tfrac{1-\rho^{4(n+1)}}{1-\rho^4} = \tfrac{\sigma_\varepsilon^2}{1-\rho^4}(1-\rho^{4(n+1)}) \end{aligned}$$

(e) When $n \to \infty$ we have

$$\begin{aligned} E[X_n] \to& 0 \\ \text{Var}[X_n] \to& \tfrac{\sigma_\varepsilon^2}{1-\rho^4} = 1 \end{aligned}$$

(f) We have

$$\begin{aligned} Y_n =&\rho X_n + \sqrt{1-\rho^2} Z_n \\ =&\rho[\rho Y_{n-1} + \sqrt{1-\rho^2} V_n] + \sqrt{1-\rho^2} Z_n \\ =&\rho^2 Y_{n-1} + \sqrt{1-\rho^2}[\rho V_n + Z_n] \end{aligned}$$

which has the same structure as for X_n and the results become identical.

(g) We have

$$\begin{aligned}E[X_nY_n] =&E[X_n(\rho X_n + \sqrt{1-\rho^2}Z_n)]\\ =&\rho E[X_n^2]\\ =&\rho[1-\rho^{4n}+\rho^{4n}\mu_0^2]\\ =&\rho+\rho^{4n+1}(\mu_0^2-1)\end{aligned}$$

(h) We see that $E[X_nY_n] \to \rho$

(i) We then see that the limit distribution for $X_n, Y_n)$ indeed is the target distribution.

We see that the convergence speed is geometric in ρ^2 for the mean and geometric in ρ^4 for the variances and the correlations.

Exercise 4 (a) The general idea is to simulate $(\phi, x_1, ..., x_t)$ by a proposal distribution $q_\phi(\phi)q_1(x_1|\phi)\prod_{i=2}^t q_i(x_i|x_{i-1},\phi)$ and then use the importance sampling technique to get importance weights

$$\begin{aligned}w_t =&\frac{p(\phi, x_1, ..., x_t|y_1, ..., y_t)}{q(\phi, x_1, ..., x_t)}\\ \propto&\frac{p(\phi)p(x_1, ..., x_t|\phi)p(y_1, ..., y_t|x_1, ..., x_t, \phi)}{q(\phi, x_1, ..., x_t)}\\ =&\frac{p(\phi)p(x_1|\phi)\prod_{i=2}^t p(x_i|x_{i-1},\phi)\prod_{i=1}^t p(y_t|x_t)}{q_\phi(\phi)q_1(x_1|\phi)\prod_{i=2}^t q_i(x_i|x_{i-1},\phi)}\\ \propto& w_{t-1}\frac{p(x_t|x_{t-1},\phi)p(y_t|x_t}{q(x_t|x_{t-1},\phi)}\end{aligned}$$

showing that the weights can be calculated recursively.

Due to that the variance of the weights will increase with t, a degeneracy problem occur. This can be fixed by performing resampling at each step (or when the efficient sample size is small).

(b) The resampling step will result in that fewer and fewer unique values of ϕ will occur, in the end only one. This causes problems in estimation of ϕ due to that we then effectively only have one sample for describing the whole distribution of ϕ.

(c) Assume you have a properly weighted sample $\{(x_t^i, \boldsymbol{S}_t^i, w_t^i), i = 1, ..., M\}$ with respect to $p(x_t, \boldsymbol{S}_t|y_1, ..., y_t)$ where $\boldsymbol{S}_t^i$ are the sufficient statistics needed for calculating the distribution $p(\phi|x_1^i, ..., x_t^i, y_1, ..., y_t)$. The idea is then to update to a properly weighted sample $\{(x_{t+1}^i, \boldsymbol{S}_{t+1}^i, w_t^i), i = 1, ..., M\}$ with respect to $p(x_{t+1}, \boldsymbol{S}_{t+1}|y_1, ..., y_{t+1})$.

We have

$$\begin{aligned}p(x_t, S_t|\boldsymbol{y}_{1:t-1}) =& \int_{x_{t-1}} p(x_t, S_t|x_{t-1}, S_{t-1})p(x_{t-1}, S_{t-1}|\boldsymbol{y}_{1:t-1})dx_{t-1}dS_{t-1}\\ \approx& \sum_{i=1}^{N} w_{t-1}^i p(x_t, S_t|x_{t-1}^i, S_{t-1}^i)\\ p(x_t, S_t|\boldsymbol{y}_{1:t}) \approx& c \cdot \sum_{i=1}^{N} w_{t-1}^i p(x_t, S_t|x_{t-1}^i, S_{t-1}^i)p(y_t|x_t).\end{aligned}$$

Simulation from $p(x_t, S_t|x_{t-1}^i, S_{t-1}^i)$ (possible proposal function)

(a) Simulate $\theta^i \sim p(\theta|x_{t-1}^i, S_{t-1}^i) = p(\theta|S_{t-1}^i)$.

(b) Simulate $x_t^i \sim p(x_t|x_{t-1}^i, \theta^i)$.

(c) Update sufficient statistics

(d) By turning the simulation from the static parameter ϕ to the random variable $\boldsymbol{S}_t$, we reduce the degeneracy problem and obtain a more reliable description of the distribution for x_t and ϕ as well.

In order to estimate ϕ, we can use Rao-Blackwellization in that

$$E[\phi|y_1, ..., y_t] = E[E[\phi|\boldsymbol{S}_t]|y_1, ..., y_t]$$

where we now have explicit solutions for the inner expectation.

Exercise 5 (Weight loss programme) (a) Since $\boldsymbol{\beta}$ is only involved in the last term, we get this result directly,

We have that

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 e^{-\beta_2 x_i})^2$$

Putting this to zero, we obtain

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 e^{-\beta_2 x_i})^2.$$

One can also show that the second derivative becomes positive, showing that it is a maximum point. This shows that for given $\hat{\boldsymbol{\beta}}$ we have an explicit solution for $\hat{\sigma}^2$.

(b) Assume one wants to minimize $g(\boldsymbol{\theta})$. Newton's method:

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta} - [g''(\boldsymbol{\theta}^t)]^{-1}g'(\boldsymbol{\theta}^t).$$

In this case $\boldsymbol{\theta} = \boldsymbol{\beta}$ and $g(\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 e^{-\beta_2 x_i})^2$. We have

$$
\begin{aligned}
\frac{\partial g(\boldsymbol{\beta})}{\partial \beta_0} =& -2\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 e^{-\beta_2 x_i})\\
\frac{\partial g(\boldsymbol{\beta})}{\partial \beta_1} =& -2\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 e^{-\beta_2 x_i})e^{-\beta_2 x_i}\\
\frac{\partial g(\boldsymbol{\beta})}{\partial \beta_2} =& 2\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 e^{-\beta_2 x_i})\beta_1 e^{-\beta_2 x_i}x_i\\
\frac{\partial^2 g(\boldsymbol{\beta})}{\partial \beta_0\partial \beta_0} =& 2n\\
\frac{\partial^2 g(\boldsymbol{\beta})}{\partial \beta_0\partial \beta_1} =& 2\sum_{i=1}^{n}e^{-\beta_2 x_i}\\
\frac{\partial^2 g(\boldsymbol{\beta})}{\partial \beta_0\partial \beta_2} =& 2\beta_1\sum_{i=1}^{n}e^{-\beta_2 x_i}x_i\\
\frac{\partial^2 g(\boldsymbol{\beta})}{\partial \beta_1\partial \beta_1} =& 2\sum_{i=1}^{n}e^{-2\beta_2 x_i}\\
\frac{\partial^2 g(\boldsymbol{\beta})}{\partial \beta_1\partial \beta_2} =& 2\sum_{i=1}^{n}(y_i-\beta_0-2\beta_1 e^{-\beta_2 x_i})e^{-\beta_2 x_i}x_i\\
\frac{\partial^2 g(\boldsymbol{\beta})}{\partial \beta_2\partial \beta_2} =& -2\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 e^{-\beta_2 x_i})\beta_1 e^{-\beta_2 x_i}x_i^2+\sum_{i=1}^{n}\beta_1^2 e^{-2\beta_2 x_i}x_i^2
\end{aligned}
$$

(c) For the Fisher scoring algorithm, we replace the matrix of second derivatives with their expectation. This guarantees that the matrix becomes positive (semi-)definite due to that is the variance of the scoring function.

(d) For given β_2, we can define $z_i = e^{-\beta_2 x_i}$ and we then have an ordinary linear regression model with z_i as explanatory variable. We can then use the general results from linear regression.

This means that we can reduce the optimization down to just one variable, simplifying the problem significantly.

Exercise 6 (a) We have

$$
\begin{aligned}
\Pr(Y_i = y) =& \sum_{k=1}^{K}\Pr(C_i = k)\Pr(Y_i = y|C_i = k)\\
=& \sum_{k=1}^{K}\pi_k\frac{\lambda_k^y e^{-\lambda_k}}{y!}
\end{aligned}
$$

giving

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

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 the average of the other values. The worst value is then updated to a better (best?) value along this line. The algorithm is performing these steps iteratively until some stopping criterion is achieved. This method does not need the derivatives.

(b) We have that the complete likelihood is given by

$$\begin{aligned}
L_c(\boldsymbol{\theta}) =& \prod_{i=1}^n \pi_{c_i}\frac{\lambda_{c_i}^{y_i}e^{-\lambda_{c_i}}}{y_i!}\\
\ell_c(\boldsymbol{\theta}) =& \log L_c(\boldsymbol{\theta})\\
=& \sum_{i=1}^n [\log \pi_{c_i} + y_i\log\lambda_{c_i} - \lambda_{c_i} - \log y_i!]\\
=& \sum_{i=1}^n\sum_{k=1}^K I(c_i = k)[\log \pi_k + y_i\log\lambda_k - \lambda_k - \log y_i!]\\
Q(\boldsymbol{\theta},\boldsymbol{\theta}^t) =& E[\ell_c(\boldsymbol{\theta})|\boldsymbol{\theta}^t]\\
=& \sum_{i=1}^n\sum_{k=1}^K \Pr(C_i = k|\boldsymbol{\theta}^t)[\log \pi_k + y_i\log\lambda_k - \lambda_k - \log y_i!]\\
Q_{lagr}(\boldsymbol{\theta},\boldsymbol{\theta}^t) =& Q(\boldsymbol{\theta},\boldsymbol{\theta}^t) + \delta(\sum_{k=1}\pi_k - 1)\\
\frac{\partial}{\partial\pi_k}Q_{lagr}(\boldsymbol{\theta},\boldsymbol{\theta}^t) =& \sum_{i=1}^n \Pr(C_i = k|\boldsymbol{\theta}^t)\frac{1}{\pi_k} - \delta
\end{aligned}$$

giving

$$\begin{aligned}
\pi_k^t =& \delta^{-1}\sum_{i=1}^n \Pr(C_i = k|\boldsymbol{\theta}^t)\\
=& \frac{1}{n}\sum_{i=1}^n \Pr(C_i = k|\boldsymbol{\theta}^t)\\
\frac{\partial}{\partial\lambda_k}Q_{lagr}(\boldsymbol{\theta},\boldsymbol{\theta}^t) =& \sum_{i=1}^n \Pr(C_i = k|\boldsymbol{\theta}^t)[\frac{y_i}{\lambda_k} - 1]
\end{aligned}$$

giving

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

where the probabilities $\Pr(C_i = k|\boldsymbol{\theta}^t)$ are based on the parameter values from the previous iteration.

The EM-algorithm has the property that the (log-)likelihood values will never decrease from one iteration to another, which the plot demonstrate.

(c) A problem in this case is that the different classes are difficult to distinguish from the data, making several configurations of mixtures of Poisson data possible.

Exercise 7 (a) In simulated annealing, first a neighborhood structure is chosen defining possible changes at each iteration. Thereafter a possible proposal $\boldsymbol{\gamma}^*$ is drawn from the neighborhood of the current value $\boldsymbol{\gamma}^t$. The proposal is then accepted with a probability

$$\min[1, \exp\{[J(\boldsymbol{\gamma}^t) - J(\boldsymbol{\gamma}^*)]/\tau_t\}$$

where τ_t is the temperature at iteration t. In order to guarantee convergence to the global maximum τ_t chould convergence to zero as $c/\log(1+t)$ where c is the depth (the smalles increase needed to escape from a local minimum). In practice this leads to much too slow convergence and a faster decrease is typically used.

(b) If the temperature is chosen to be very large, just random changes are made, not using $J(\boldsymbol{\gamma})$ at all.

If a a very low temperature is chosen, changes are only made if a better proposal is found, corresponding to a greedy algorithm.

If the temperature is fixed, we obtain a Metropolis-Hastings algorithm with $J(\boldsymbol{\gamma})/\tau$ as target distribution. For $\tau = 1$ this then corresponds to a Bayesian posterior with the penalty term serving as a prior.

Exercise 8 (a) The mean field approximation approximates $p(\mu_1, \mu_2|\boldsymbol{x})$ by

$$q(\mu_1, \mu_2) = q_1(\mu_1)q_2(\mu_2)$$

that is it assumes independence between μ_1 and μ_2. Typically, for continuous variables, one uses the Gaussian distributions for each component, that is $q(\mu_j) = N(a_j, b_j^2)$ for $j = 1, 2$ and $\{(a_j, b_j^2), j = 1, 2\}$ then needs to be specified. This is typically done by minimizing the Kullback-Leibler distance between $q(\mu_1, \mu_2)$ and $p(\mu_1, \mu_2|\boldsymbol{x})$. This changes the integration problem to an optimization problem.

(b) Note that if the class-membership was known we have $p(\mu_1, \mu_2|\boldsymbol{x}, \boldsymbol{c}) = p(\mu_1|\boldsymbol{x}, \boldsymbol{c})p(\mu_2|\boldsymbol{x}, \boldsymbol{c})$, that is they are independent.

For the last row, the two classes are quite separated, making it relatively easy to identify which x_i's that belong to the two classes. In that case, the mean field approximation will become quite good. As there becomes more ucnertainty to which classes that the x_i's belong to, the mean field approximation becomes worse due to that there will be more dependence between μ_1 and μ_2.