UiO : Matematisk institutt
Det matematisk-naturvitenskapelige fakultet

STK-4051/9051 Computational Statistics Spring 2024
Comments to exercise 8

Instructor: Odd Kolbjørnsen, oddkol@math.uio.no

Exercise 26 (Sequential importance sampling)
In population ecology, variations of the population sizes for a specific animal is measured through time-series observations on the number of animals caught in traps. Assume y_t is the number of animals caught at time t (the time-scale is typically in years).

A simple model in this case (defining x_t to be the logarithm of the population size) is

$$
\begin{aligned}
x_1 &\sim N(\mu, \sigma^2/(1-a^2)) \\
x_t &\sim N(\mu + a(x_{t-1} - \mu), \sigma^2) \\
y_t &\sim \text{Poisson}(\exp\{x_t\})
\end{aligned}
$$

The following data (which is also given on the web-page under the name ***sim_animal_trap.txt***) are data simulated from the model above using $\mu = 2, a = 0.9, \sigma = 0.5$. The first row corresponds to the first 18 time-points and so on.

2	7	8	4	7	7	7	8	11	10	8	4	8	9	19	12	35	39
14	5	6	5	6	1	0	2	1	3	6	4	1	0	2	3	1	2
0	1	3	0	3	9	4	13	23	15	7	9	10	6	3	12	16	29
28	18	13	6	8	14	25	14	17	11	19	39	55	71	83	61	60	44
57	26	24	47	20	53	65	68	56	48	26	23	29	17	2	30	24	52
27	20	13	13	18	19	5	4	10	9								

(a). Show that $x_t \sim N(\mu, \sigma^2/(1-a^2))$ for all t. Discuss this property.

Solution to exercise 26. (a). Assume $x_{t-1} \sim N(\mu, \sigma^2/(1-a^2))$. Then

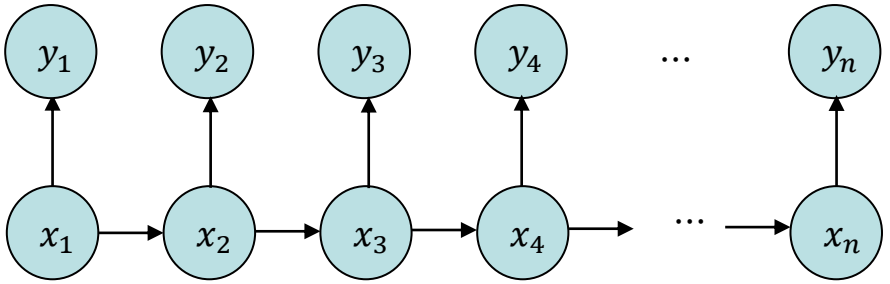
$$x_t = \mu + a(x_{t-1} - \mu) + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma^2)$$

were ε_t is independent of x_{t-1}. Then x_t is a lineary combination of Gaussian variables, making itself Gaussian. Further

$$
\begin{aligned}
E[x_t] &= \mu + a(E[x_{t-1} - \mu) + E[\varepsilon_t] = \mu \\
\text{Var}[x_t] &= a^2 \text{Var}[x_{t-1}] + \text{Var}[\varepsilon_t] = a^2 \tfrac{\sigma^2}{1-a^2} + \sigma^2 = \tfrac{\sigma^2}{1-a^2}
\end{aligned}
$$

#26

(b). Write down the posterior (or conditional) distribution for $\boldsymbol{x}_{1:T} = (x_1, ..., x_T)$ given $\boldsymbol{y}_{1:T} = (y_1, ..., y_T)$ (up to a proportionality constant).



(b). We have

$$p(\boldsymbol{x}_{1:T}|\boldsymbol{y}_{1:T}) = \frac{p(\boldsymbol{x}_{1:T})p(\boldsymbol{y}_{1:T}|\boldsymbol{x}_{1:T})}{p(\boldsymbol{y}_{1:T})}$$

$$\propto p(x_1)p(y_1|x_1)\prod_{t=2}^{T} p(x_t|x_{t-1})p(y_t|x_t)$$

were each of the densities involved are specified through the given model.

(c). Consider first a case where $\boldsymbol{x}_{1:T}$ is sampled from the prior, that is

$$g(\boldsymbol{x}_{1:t}) = g_1(x_1)\prod_{s=2}^{t} g_s(x_s|x_{s-1})$$

with $g_1(x_1) = N(\mu, \sigma^2/(1-a^2))$ and $g_s(x_s|x_{s-1}) = N(\mu + a(x_{s-1}-\mu), \sigma^2)$. Calculate the importance weight in this case and show that it can be written recursively as

$$w_t(\boldsymbol{x}_{1:t}) = w_{t-1}(\boldsymbol{x}_{1:t-1})u_t(y_t, x_t)$$

for properly defined functions $w_1(x_1)$ and $u_t(y_t, x_t)$. (Here $\boldsymbol{x}_{1:t} = (x_1, ..., x_t)$.)

(c). We have then that $g(\boldsymbol{x}_{1:t}) = p(\boldsymbol{x}_{1:t})$ so that

$$w_t(\boldsymbol{x}_{1:t}) = \frac{p(\boldsymbol{x}_{1:t}|\boldsymbol{y}_{1:t})}{g(\boldsymbol{x}_{1:t})} \propto p(\boldsymbol{y}_{1:t}|\boldsymbol{x}_{1:t}) = p(\boldsymbol{y}_{1:t-1}|\boldsymbol{x}_{1:t-1})p(y_t|x_t) \propto w_{t-1}(\boldsymbol{x}_{1:t-1})p(y_t|x_t)$$

We only need the weights up to a proportionality constant (since we will normalize them anyway), showing the result with $u_t(y_1, x_t) = p(y_t|x_t)$.

(*d*). Implement a Sequential Monte Carlo algorithm based on the previously results and try it out on the data given above. Use $N = 10\,000$ and resampling at each time-point.

For each t, estimate $\hat{x}_{t|t} = E[x_t|\boldsymbol{y}_{1:t}]$ and also estimate the 0.025 and 0.975 quantiles in the distribution $p(x_t|\boldsymbol{y}_{1:t})$. Plot these estimates and quantiles in the same plot.

Filtering

Also calculate the effective sample size just before you do resampling. Plot this as a function of time.

Hint: Modify one of the R-scripts with Sequential Monte Carlo from the web-page.

(*e*). Assume our interest now is on $\hat{x}_{t|T} = E[x_t|\boldsymbol{y}_{1:T}]$, that is the state estimates based on all data. Explain how these estimates can be obtained from your Sequential Monte Carlo algorithm. Plot these estimates (together with 0.025 and 0.975 quantiles in the distribution $p(x_t|\boldsymbol{y}_{1:T})$ on top on those you plotted in (*d*). Discuss similarities and differences.

Smoothing

Also look at the number of unique values that the approximation of $\hat{x}_{t|T} = E[x_t|\boldsymbol{y}_{1:T}]$ is based on as a function of time.

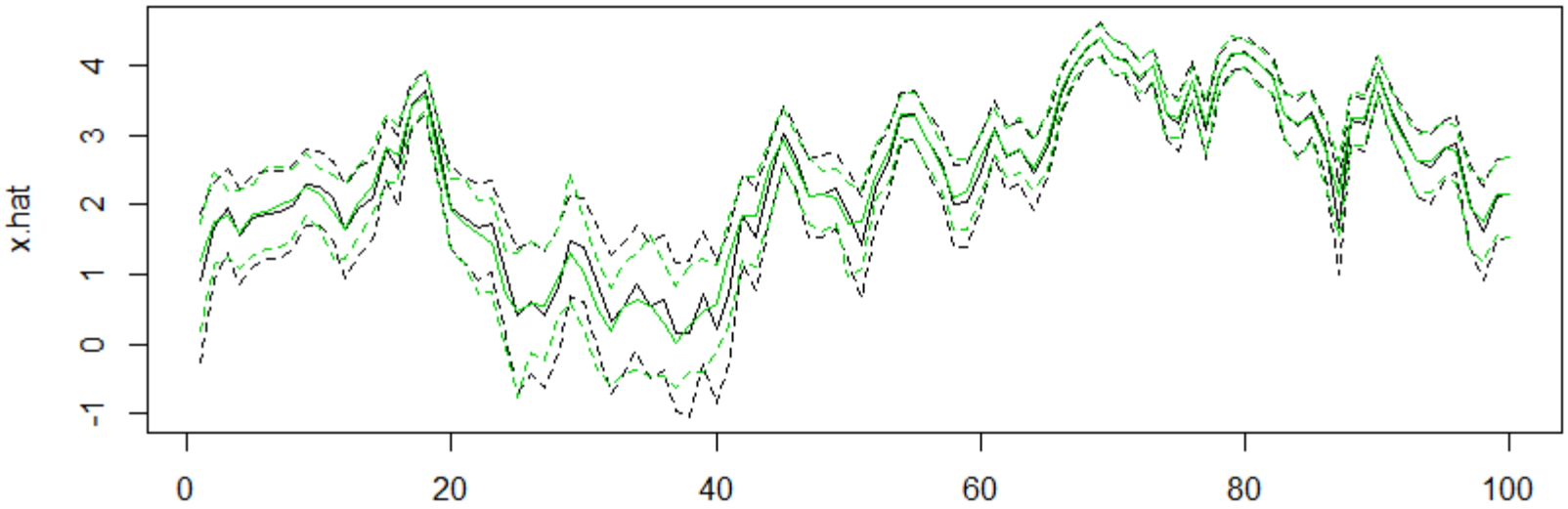
(*e*). The importance weights we calculate is based on the whole sequence $\boldsymbol{x}_{1:t}$. Therefore the samples $(\boldsymbol{x}^i_{1:t}, w^i_t)$ are properly weighted with respect to $p(\boldsymbol{x}_{1:t}|\boldsymbol{y}_{1:t})$. When $t = T$, we then obtain properly weighted samples from $p(\boldsymbol{x}_{1:T}|\boldsymbol{y}_{1:T})$.

When we perform resampling in the algorithm, note that we then need to resample the whole sequence $\boldsymbol{x}_{1:t}$.

A problem when looking at these "smoothed" estimates is that for t small, the number of unique samples is very low. Imn this case, were T is not too large and the number of samples N is large enough, we do however still get reasonable estmates and uncertainty measures even for x_1.

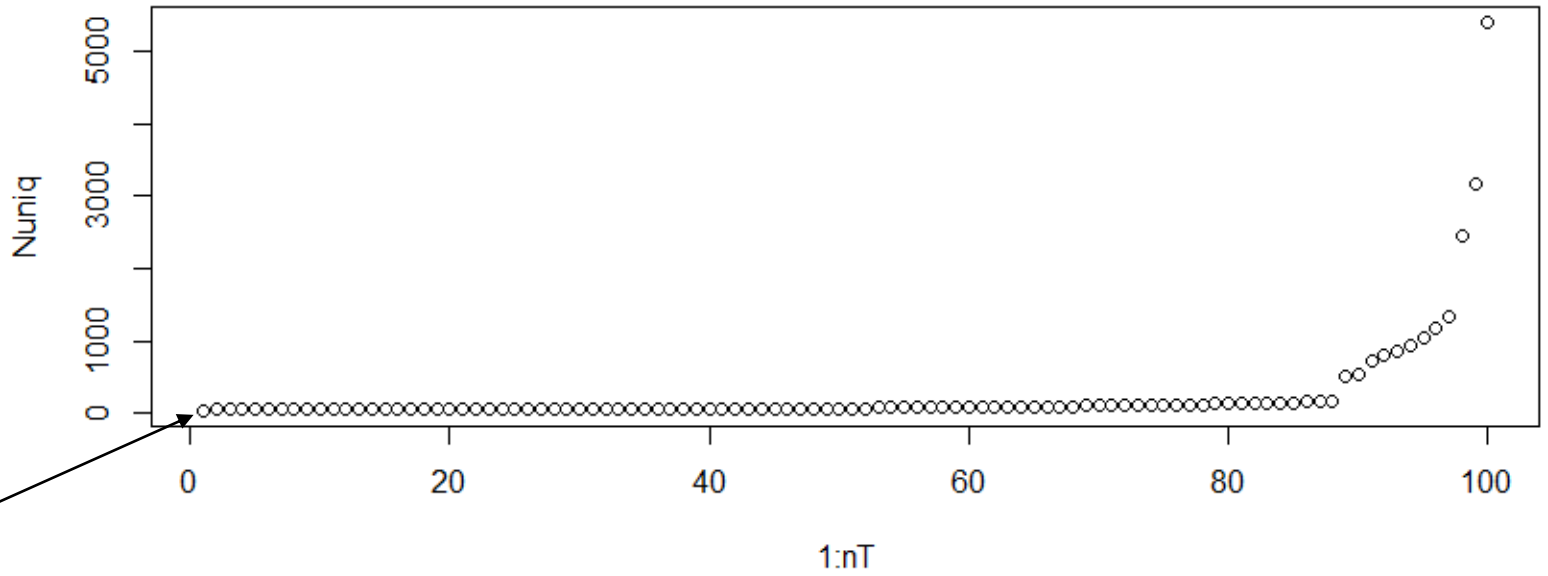
```
for(i in 2:nT)
{
  x[i,] = rnorm(N,mu+a*(x[i-1,]-mu),sigma)
  w = w*dpois(y[i],exp(x[i,]))
  w = w/sum(w)
  N.eff[i] = N/sum(w^2)
  #Resample
  ind = sample(1:N,N,replace=T,prob=w)
  x[1:i,] = x[1:i,ind]  # NB resample entire path
  w= rep(1/N,N)
  x.hat[i,1] = mean(x[i,])
  x.hat[i,2:3] = quantile(x[i,],c(0.025,0.975))
}
```

Filtering: Computed on the fly



```
x.smo = matrix(nrow=nT,ncol=3)
N.unique = rep(NA,nT)
for(i in 1:nT)
{
  x.smo[i,1] = mean(x[i,])
  x.smo[i,2:3] = quantile(x[i,],c(0.025,0.975))
  cat("Unique",i,length(unique(x[i,])),"\n")
  N.unique[i] = length(unique(x[i,]))
}
```

Smoothing
Computed
after loop
Is finished



49

Ex 26

a) See solution, it is nice to know that the distribution is stationary.

b) Solution is a bit brief, but see also 27a) to get the expression for the likelihood

c) Too ease the understanding of the derivation, think of this as a proof by induction, the lines here is the general step from t-1 to t. Use also the expression in 26b to derive the result

d) See code,

e) The comment in the solution, is good.

Exercise 27 (Improvements of SIS)
We will in this exercise consider the same problem as the one in exercise 26, but now see if we are able to imrpove the algorithm described there by using better proposal distributions. The main idea is that the simple proposal used in exercise 26 do not take the data into account at all, and that using the data should help us in simulating more reasonable x's.

(a). Consider first a simple situation where $x \sim N(\mu, \sigma^2)$ and $y \sim \text{Poisson}(\exp\{x\})$.

Write down the posterior distribution for x given y, $p(x|y)$ (up to a proportionality constant).

Solution to exercise 27. (a). We have

$$\begin{aligned} p(x|y) &\propto p(x)p(y|x) \\ &\propto \exp\left(-\tfrac{1}{2\sigma^2}(x-\mu)^2\right)\frac{\exp(x)^y \exp(-\exp(x))}{y!} \\ &\propto \exp\left(-\tfrac{1}{2\sigma^2}(x^2-2\mu x) + yx - \exp(x)\right) \end{aligned}$$

(*b*). Consider the logarithm of $p(x|y)$, and assume we want to approximate this by a function of the form Const $-\frac{1}{2\tilde{\sigma}^2}(x-\tilde{\mu})^2$. What kind of distribution does this correspond to?

Argue why a reasonable approximation for e^x is

$$e^{\mu} + e^{\mu}(x-\mu) + \tfrac{1}{2}e^{\mu}(x-\mu)^2$$

in this case. Use this approximation to derive $\tilde{\mu}$ and $\tilde{\sigma}^2$.

$$\exp\left(-\tfrac{1}{2\sigma^2}(x^2 - 2\mu x) + yx - \exp(x)\right)$$

(*b*). We have that this approximation corresponds to a Gaussian approximation to $p(z|y)$.

From the prior we have that x should not be too far from μ. The approximation of $\exp(x)$ corresponds to a Taylor approximation for $\exp(\mu)$ around μ. We then get

$$\begin{aligned}
\log p(x|y) \approx & \text{Const} - \tfrac{1}{2\sigma^2}(x^2 - 2\mu x) + yx - e^{\mu} - e^{\mu}(x-\mu) - \tfrac{1}{2}e^{\mu}(x-\mu)^2 \\
= & \text{Const} - (\tfrac{1}{2\sigma^2} + \tfrac{1}{2}e^{\mu})x^2 + (\tfrac{\mu}{\sigma^2} + y - e^{\mu} + \mu e^{\mu})x \\
= & \text{Const} - (\tfrac{1}{2\sigma^2} + \tfrac{1}{2}e^{\mu})[x - \frac{\frac{\mu}{\sigma^2} + y - e^{\mu} + \mu e^{\mu}}{\frac{1}{\sigma^2} + e^{\mu}}]^2 \\
= & \text{Const} - \frac{1}{2\tilde{\sigma}^2}(x-\tilde{\mu})^2
\end{aligned}$$

with

$$\begin{aligned}
\tilde{\sigma}^2 = & \frac{1}{\frac{1}{\sigma^2} + e^{\mu}} = \frac{\sigma^2}{1 + \sigma^2 e^{\mu}} \\
\tilde{\mu} = & \frac{\frac{\mu}{\sigma^2} + y - e^{\mu} + \mu e^{\mu}}{\frac{1}{\sigma^2} + e^{\mu}} = \frac{\mu + \sigma^2(y - e^{\mu} + \mu e^{\mu})}{1 + \sigma^2 e^{\mu}}
\end{aligned}$$

(c). Consider now the setting of exercise 26. Use the approximation above to suggest a proposal distribution for x_t that is approximately $p(x_t|x_{t-1}, y_t)$.

Modify your algorithm to consider this proposal and run it on the given data.

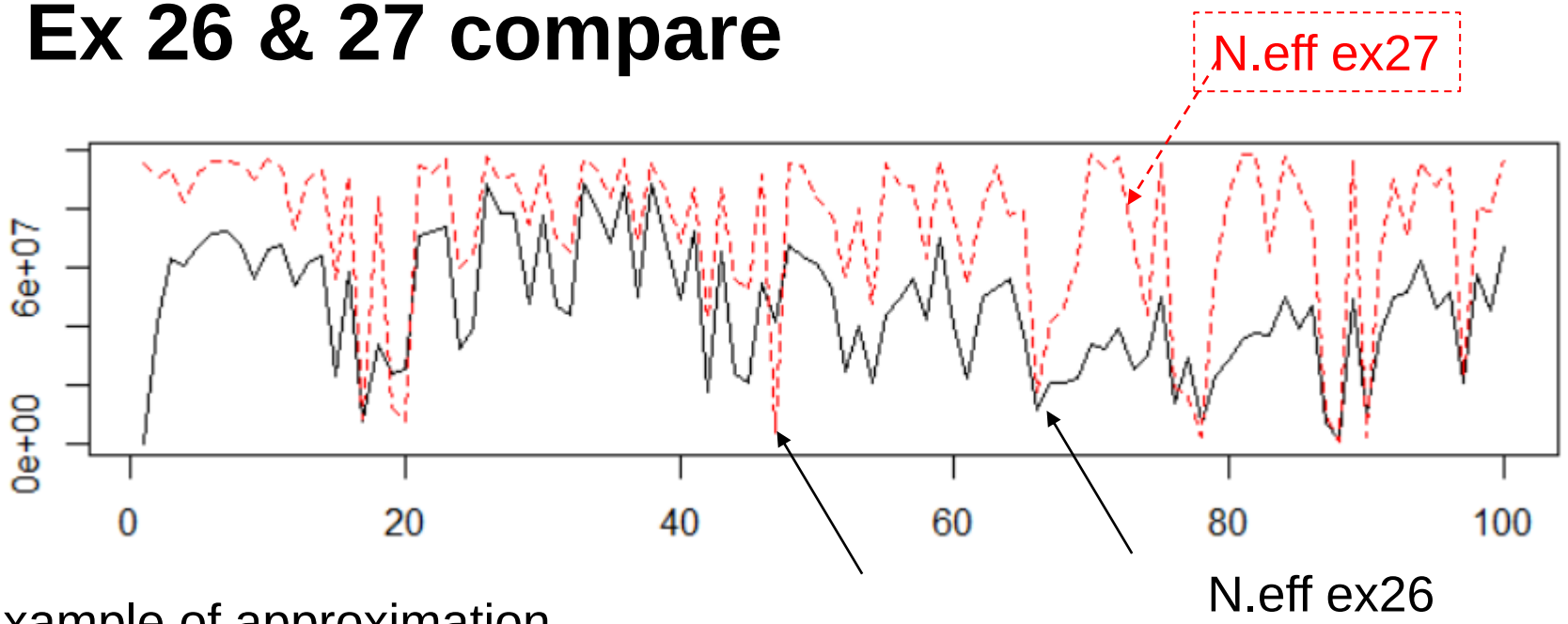
(d). Use some measures to evaluate the performance of this modification compared to the simpler algorithm used in exercise 26. Which algorithm do you prefer?

Make a proposal from a normal distribution using the approximation

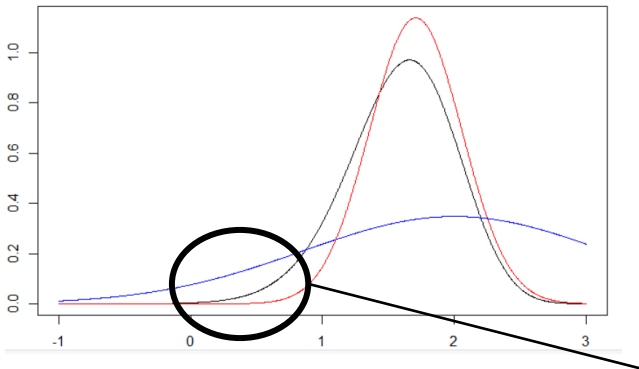
$$\tilde{\sigma}^2 = \frac{1}{\frac{1}{\sigma^2} + e^{\mu}} = \frac{\sigma^2}{1 + \sigma^2 e^{\mu}}$$

$$\tilde{\mu} = \frac{\frac{\mu}{\sigma^2} + y - e^{\mu} + \mu e^{\mu}}{\frac{1}{\sigma^2} + e^{\mu}} = \frac{\mu + \sigma^2(y - e^{\mu} + \mu e^{\mu})}{1 + \sigma^2 e^{\mu}}$$

Ex 26 & 27 compare



Example of approximation



True distribution

Gaussian approx

Prior

The approximation (ex 27) is not allways better

The main problem is the approximation in the left tail. In this region Gauss approx. is too low which gives large weights

Ex 27 comments

a) Since y is fixed, we can get rid of y!

b) Lots of computations here. In principle it is just a taylor expansion. The details are a bit messy….

c) It is allways a challenge to program complex expressions, check the implementation twice (or more).

d) The upside is the effective number of samples increases with more than 50% on average

Ex 29

Exercise 29 (Variance reduction)
Hammersley and Handscomb (1964) use the integration of $\phi(x) = (e^x - 1)/(e - 1)$ on $(0, 1)$ as a test problem of variance reduction techniques. Achieve as large a variance reduction as you can compared to the naive Monte Carlo integration based on uniform sampling on $(0, 1)$. (Hammersley and Handscomb achieved 4 million).

- This is for you to get a feeling with the different methods. So not much to say here except that you should try it out.

- Note on the lambda for control variates:
  - We can compute this ratio using the input variables:
  - That is compute the variance and covariance of: $h(X_i)$ and $c(Y_i)$
  - $\lambda = -\text{cov}(h(X), c(Y))/\text{var}(c(Y))$
  - Remember that this number was quite robust towards small deviations

$$\hat{\mu}_{MC} = N^{-1} \sum_{i=1}^{N} h(\mathbf{X}_i)$$

$$\hat{\theta}_{MC} = N^{-1} \sum_{i=1}^{N} c(\mathbf{Y}_i)$$

$$\lambda = -\frac{\text{cov}[\hat{\mu}_{MC}, \hat{\theta}_{MC}]}{\text{var}[\hat{\theta}_{MC}]}$$

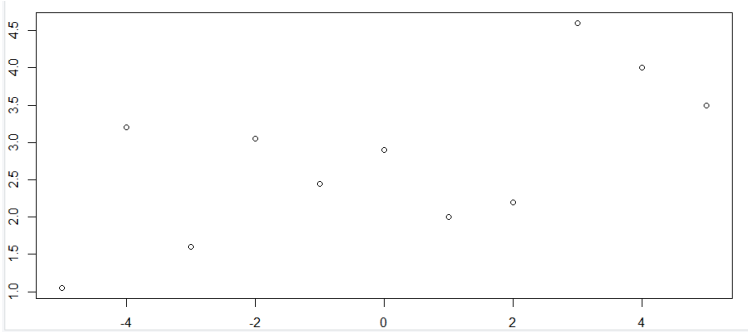
Exercise 32 (The effect of model selection)
The following example is adapted from Hjorth (1994). Consider the time series data given in the table.

x	-5.00	-4.00	-3.00	-2.00	-1.00	0.00	1.00	2.00	3.00	4.00	5.00
y	1.05	3.20	1.60	3.05	2.45	2.90	2.00	2.20	4.60	4.00	3.50

When you plot them (do it), it will become clear that there is considerable uncertainty as to whether there is an underlying consistent growth or not. Yet this issue is of crucial importance for forecasting. If there is evidence of such a trend, we may take the view that it is likely to continue. Consider two competing models:

$$M_0: \quad y_t = \beta_0 + \varepsilon_t, \quad t = -n, -n+1, ..., -1, 0, 1, ..., n$$
$$M_1: \quad y_t = \beta_0 + \beta_1 t + \varepsilon_t, \quad t = -n, -n+1, ..., -1, 0, 1, ..., n$$

where the ε_t's are independent errors. In a situation like this you might like to select one of the models from the empirical evidence available and use it to forecast. The problem addressed in this exercise is to what extent the model selection influence the bias and variability of the forecast. The traditional way in statistics is to ignore the issue completely and proceed with standard theory as if no data-driven selection had taken place at all. How wrong is this exactly?



As an example, consider a selection rule based on statistical significance. Let $\hat{\beta}_1$ be the least squares estimate of β, and s_1 its standard error. Proclaim M_1 if the hypothesis $H : \beta_1 = 0$ is rejected. This means that the procedure for predicting $\theta = E(y_{t_0})$ is

$$\hat{\theta} = \begin{cases} \overline{y}, & \text{if } \frac{|\hat{\beta}_1|}{s_1} < t_{\alpha/2} \\ \hat{\beta}_0 + \hat{\beta}_1 t_0, & \text{if } \frac{|\hat{\beta}_1|}{s_1} \geq t_{\alpha/2} \end{cases}$$

Here $\overline{y}$ is the mean of the observations, $(\hat{\beta}_0, \hat{\beta}_1)$ is the least squares estimates under M_1, s_1 is the estimated standard error of $\hat{\beta}_1$ and t_α is the α percentile of the t-distribution with $2n - 1$ degrees of freedom.

(a). For $t_0 = 5$, estimate $E\hat{\theta}$ and $sd(\hat{\theta})$ as a function of β_1 by running the following simulation experiment. Let $n = 5$, ε_t Gaussian distributed with $\sigma^2 = \text{var}(\varepsilon_t) = 1$, make your own choice of β_0 and let β_1 vary from, say -1.0 to 1.0 at step 0.05 (or according to another scheme if you so prefer). Let $\alpha = 0.05$. For each set of parameters, simulate data, compute $\hat{\theta}$ and repeat the number of times you find necessary. Register for each choice of β_1 in how many of the simulations model 1 was chosen.

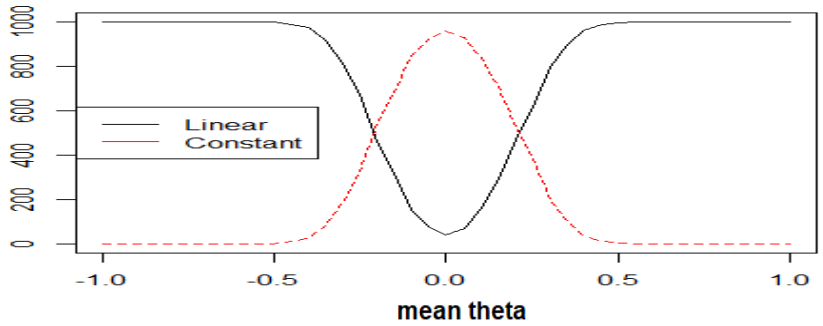
(b). Plot a density plot of θ for $\beta_1 = 0.0, 0.2, 0.3, 0.6$.

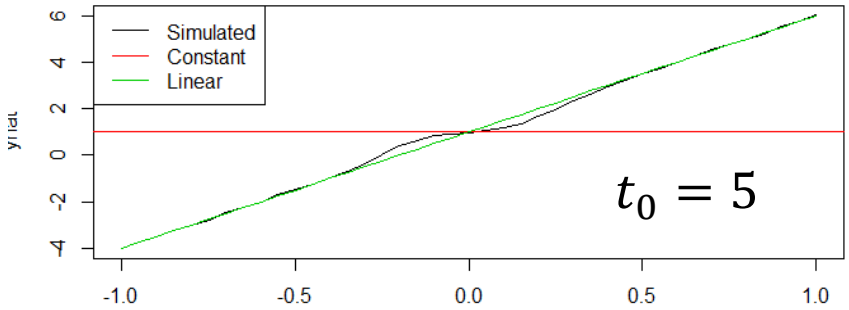
(c). Repeat (a) for $t_0 = 8$.

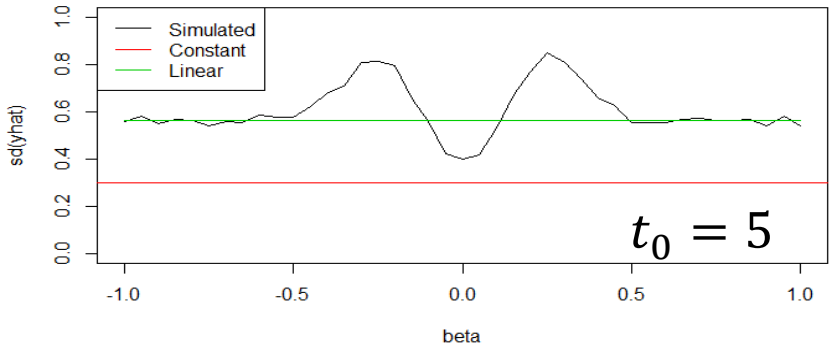
(d). Compare results in (a) and (b) with those found by standard theory that ignores the data-dependent model selection that took place.

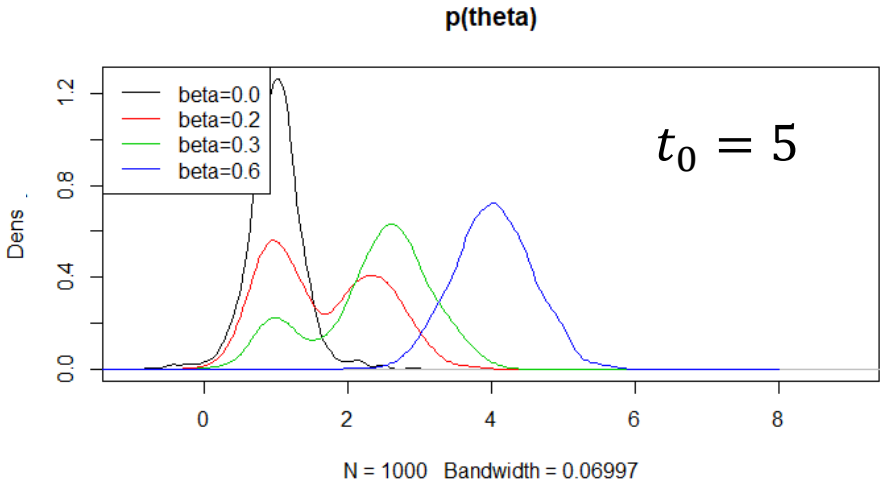
Ex 32

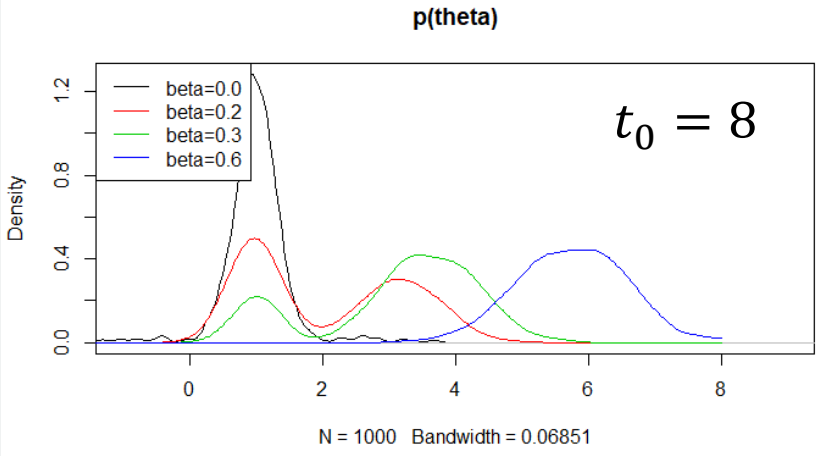
This task is just an example of how you can use simulations to check the method you are using. In this case for model selection. See code in R-file for implementation





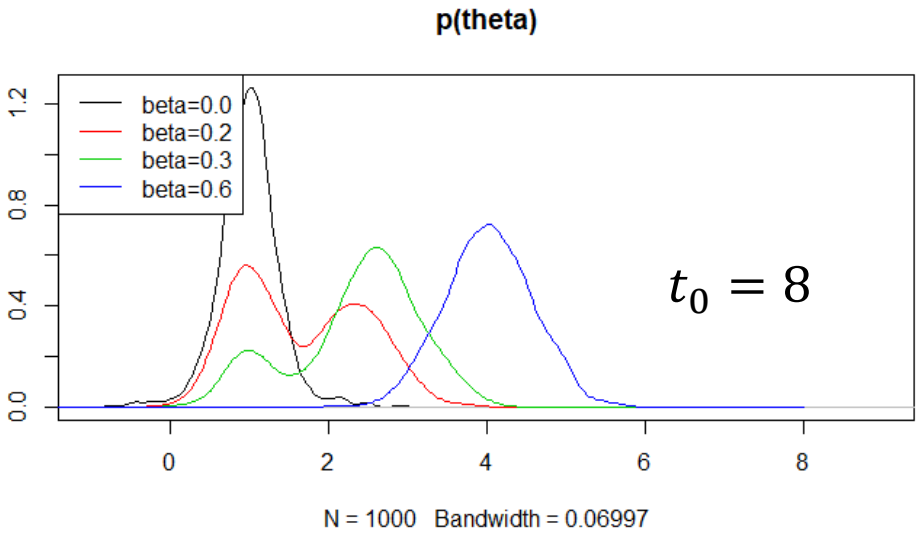






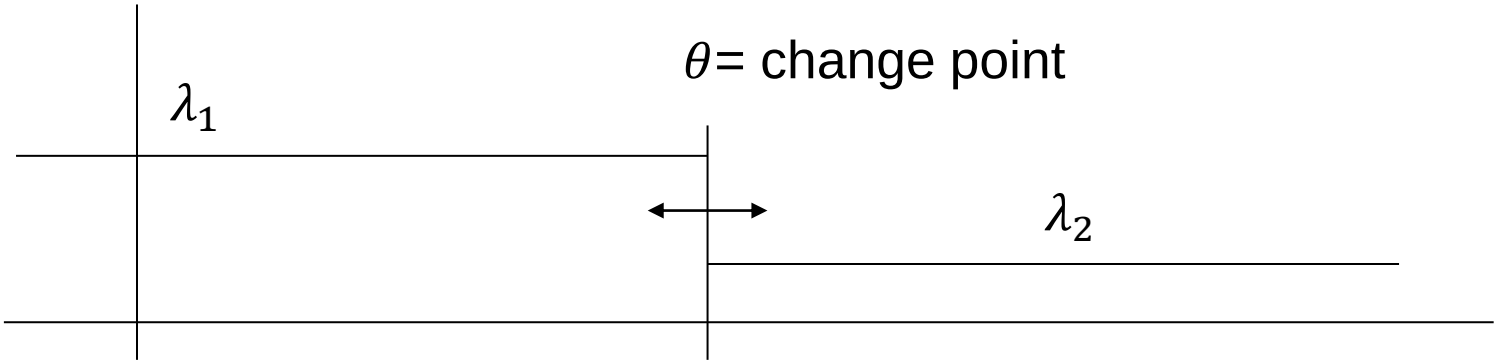
Comments

- The standard approach does not properly take the uncertainty into account when it is used for model selection.
- The true model is actually bi-modal.



6.4. Figure 6.12 shows some data on the number of coal-mining disasters per year between 1851 and 1962, available from the website for this book. These data originally appeared in [434] and were corrected in [349]. The form of the data we consider is given in [91]. Other analyses of these data include [445, 525].

The rate of accidents per year appears to decrease around 1900, so we consider a change-point model for these data. Let $j = 1$ in 1851, and index each year thereafter, so $j = 112$ in 1962. Let X_j be the number of accidents in year j, with $X_1, \ldots, X_\theta \sim$ i.i.d. Poisson(λ_1) and $X_{\theta+1}, \ldots, X_{112} \sim$ i.i.d. Poisson(λ_2). Thus the change-point occurs after the θth year in the series, where $\theta \in \{1, \ldots, 111\}$. This model has parameters θ, λ_1, and λ_2. Below are three sets of priors for a Bayesian analysis of this model. In each case, consider sampling from the priors as the first step of applying the SIR algorithm for simulating from the posterior for the model parameters. Of primary interest is inference about θ.



a. Assume a discrete uniform prior for θ on $\{1, 2, \ldots, 111\}$, and priors $\lambda_i|a_i \sim$ Gamma$(3, a_i)$ and $a_i \sim$ Gamma$(10, 10)$ independently for $i = 1, 2$. Using the SIR approach, estimate the posterior mean for θ, and provide a histogram and a credible interval for θ. Provide similar information for estimating λ_1 and λ_2. Make a scatterplot of λ_1 against λ_2 for the initial SIR sample, highlighting the points resampled at the second stage of SIR. Also report your initial and resampling sample sizes, the number of unique points and highest observed frequency in your resample, and a measure of the effective sample size for importance sampling in this case. Discuss your results.

b. Assume that $\lambda_2 = \alpha\lambda_1$. Use the same discrete uniform prior for θ and $\lambda_1|a \sim$ Gamma$(3, a)$, $a \sim$ Gamma$(10, 10)$, and $\log \alpha \sim$ Unif$(\log 1/8, \log 2)$. Provide the same results listed in part (a), and discuss your results.

c. Markov chain Monte Carlo approaches (see Chapter 7) are often applied in the analysis of these data. A set of priors that resembles the improper diffuse priors used in some such analyses is: θ having the discrete uniform prior, $\lambda_i|a_i \sim$ Gamma$(3, a_i)$, and $a_i \sim$ Unif$(0, 100)$ independently for $i = 1, 2$. Provide the same result listed in part (a), and discuss your results, including reasons why this analysis is more difficult than the previous two.

Ex 6.4

- See the code for all three examples.
  - Note that frequently we are in the situation that we have many products of numbers. And then vi divide by a product of some other numbers. In these cases. It is always recommended to working on the log-scale as this is more stable

$$\frac{\prod_{i=2}^{n} f(x_i|x_{i-1})}{\prod_{i=2}^{n} g(x_i|x_{i-1})} = \exp\left\{\sum_{i=2}^{n} \log(f(x_i|x_{i-1})) - \log(g(x_i|x_{i-1}))\right\}$$

Often give numerical problems e.g. (0/0) or Nan/Nan

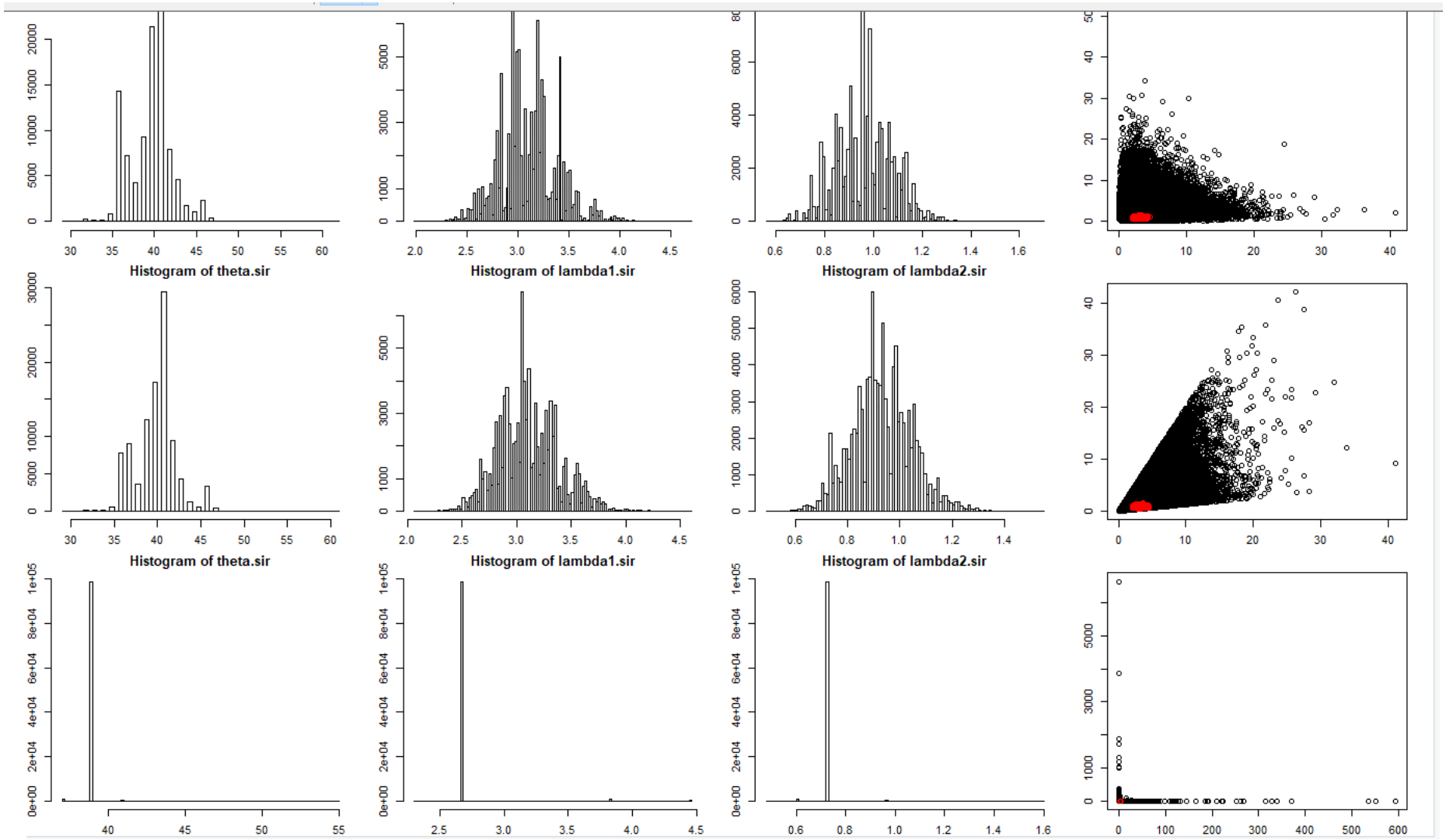
Better

```r
#Exercise 6.4
coal = read.table("V:/PROMAX/PROMAX-Odd/16_UIO/STK4051-9051/Lectures_Odd/Ex8/coal.dat",header=
coal$num = 1:112
par(mfrow=c(3,4))
set.seed(43534)
N = 100000
#a
theta = sample(1:111,N,replace=T)
a1 = rgamma(N,shape=10,rate=10)
lambda1 = rgamma(N,shape=3,rate=a1)
a2 = rgamma(N,shape=10,rate=10)
lambda2 = rgamma(N,shape=3,rate=a2)
w = rep(NA,N)
for(i in 1:N)
{
  x1 = coal$disasters[coal$num<=theta[i]]
  x2 = coal$disasters[coal$num>theta[i]]
  w[i] = sum(dpois(x1,lambda1[i],log=TRUE))+        #Working on log-scale is more stable
             sum(dpois(x2,lambda2[i],log=TRUE))
}
w = exp(w-max(w))
w = w/sum(w)
neffa=1/sum(w^2)
```

Given theta
all variables are independent

This sets the level of the weights such that the largest value before normalization is 1

Ex 6.4


Histogram of theta.sir
Histogram of lambda1.sir
Histogram of lambda2.sir
Histogram of theta.sir
Histogram of lambda1.sir
Histogram of lambda2.sir

Ex 6.4

- In the posterior of a and b
  - The marginal distributions are similar
  - Joint distribution of lambda 1 and 2 are different
- The estimates are robust towards the formulation of prior distribution

```
[1] "Prior from (a)"
Estimate of theta 39.78553
Credibility interval for theta: 36 46
Estimate of lambda1 3.105188
Credibility interval for lambda1: 2.566814 3.742627
Estimate of lambda2 0.9649861
Credibility interval for lambda2: 0.7483087 1.186242
[1] "Prior from (b)"
Estimate of theta 40.08188
Credibility interval for theta: 36 46
Estimate of lambda1 3.106418
Credibility interval for lambda1: 2.605762 3.703612
Estimate of lambda2 0.9308218
Credibility interval for lambda2: 0.7120824 1.167468
[1] "Prior from (c)"
Estimate of theta 38.99738
Credibility interval for theta: 39 39
Estimate of lambda1 2.677348
Credibility interval for lambda1: 2.662579 2.662579
Estimate of lambda2 0.7272437
Credibility interval for lambda2: 0.7268339 0.7268339
```

- In c. The method is a failure, we need many more samples to get this right
- A too wide prior is sometimes not helpful

```
Effective number of samples :
86.77606  b: 191.9267  c: 1.021563
```

Important to know when the method has failed!