# Lab 4 - Computational Statistics

*Gustav Sternelöv*

*February, 24, 2016*

## Assignment 1

### 1.1

The Metropolis-Hastings algorithm is used for generating values from the following distribution:

$$f(x) \propto x^5 e^{-x}$$

In the first step is the log-normal distribution the chosen proposal distribution and the chosen starting point is $X_0 = 2$. A time-series plot of the chain is generated to analyse the convergence of the chain and if there is a burn-in period.
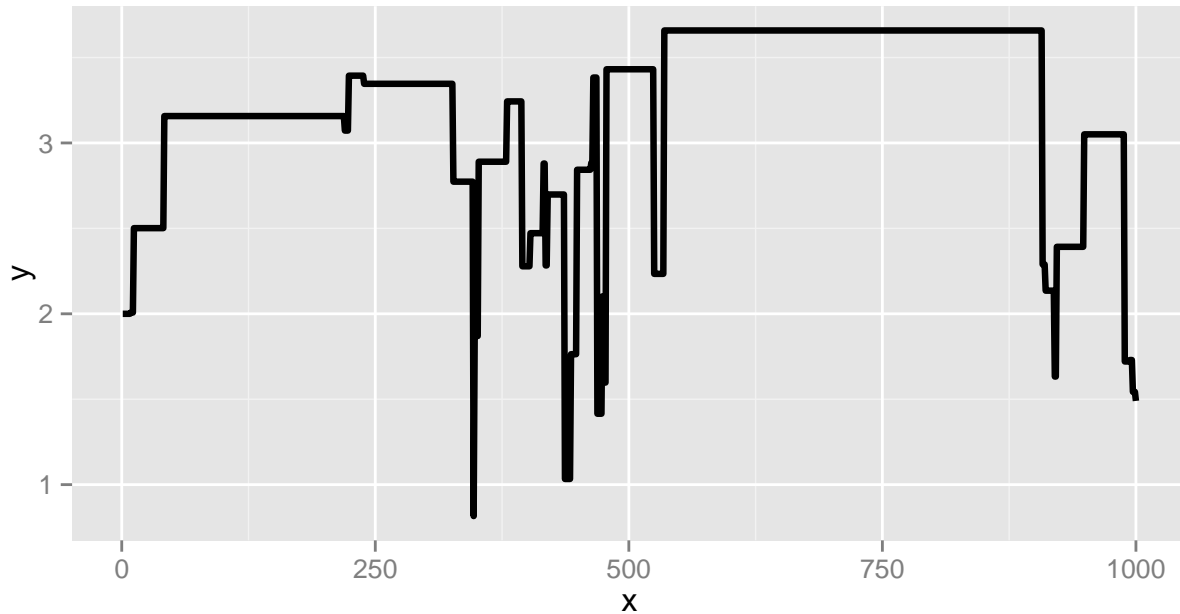


The chain never converges, the patterin is irregular over the whole time-series, and therefore there is no burn-in period to analyse.

### 1.2

In the second step is it the same distribution as in *1.1* that the algorithm is supposed to generate values from. The difference is the proposal distribution which now is a chi-square distribution, $\chi^2(floor(X_t + 1))$, where *floor(x)* means that the integer part of x is the inserted value. The starting point is $X_0 = 2$.
The chain is again plotted as a time-series in order to analyse the convergence and the eventual burn-in period.

1
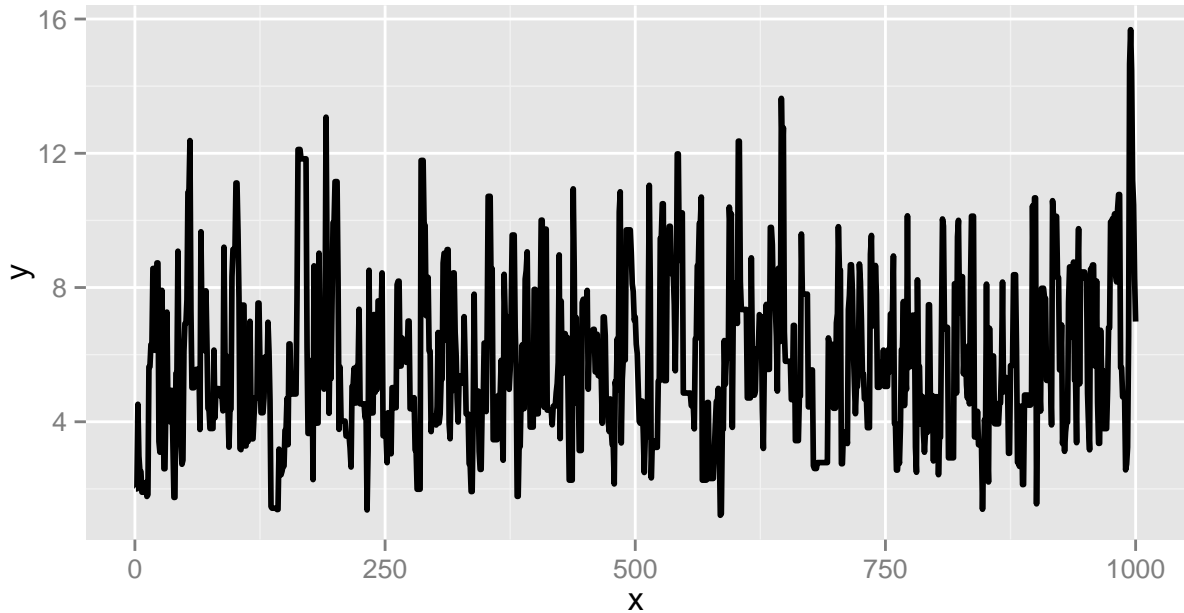
With the mentioned chi-square distribution as proposal distribution does the chain converge very quickly. Meaning that the chain has converged to the target pdf/density.

The burn-in period is not very evident by an visual eximination. Theoretically, according to Martinez(*Computational Statistics Handbook with Matlab*, 2001), the burn-in period can be assumed to be around 1-2 % of $n$ if $n$ is large enough.
In the example above $n$ is 1000 and the burn-in period would then be around 10-20 of the first values in the chain. This is thought to be reasonable as it can be seen in the plot that the chain converges very fast.

## 1.3

A comparison between the results obtained in step and step 2 shows that the chi-square distribution seem to be a better choice than the log-normal distribution for the proposal distribution.
The chain received for the log-normal as proposal distribution never converged. The pattern of the chain may suggest that a lower variance could be an option. However, the conclusion of the comparison between the results is straightforward. The chi-square distribution used converged and clearly seem to be the better choice.

Why?

Should be connected to how well the form of the respective proposal distributions fits the target distribution. It is enough for the proposal densities to be of "regular form". That means that the demand of the proposal distribution is that it has the same support with nonzero density as the target pdf. An example of how the respective density functions looks:
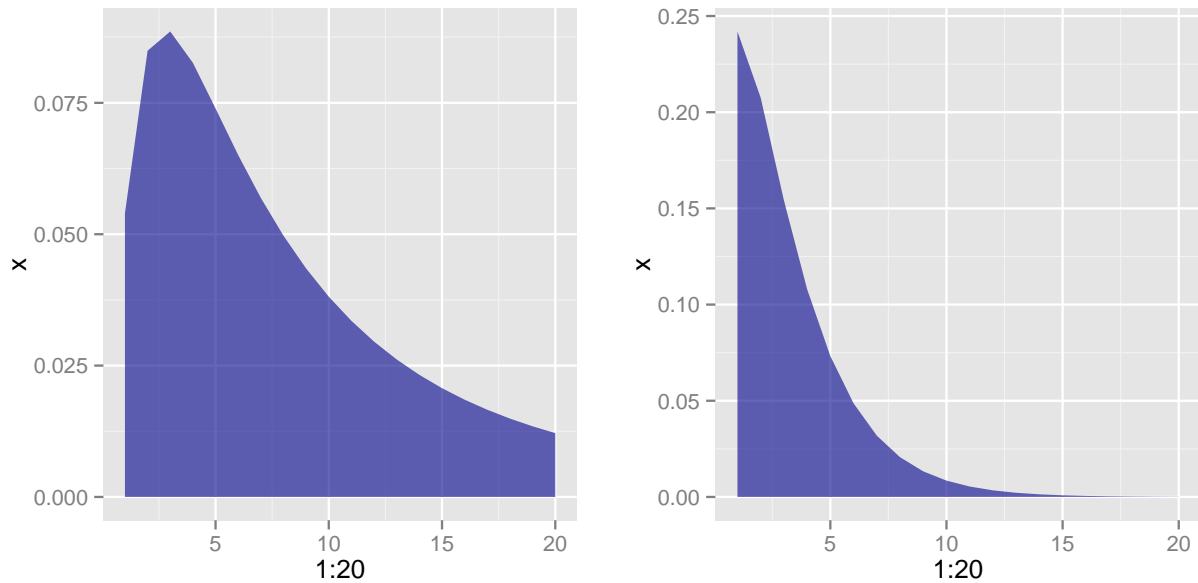
More things that can have an impact on the properness for an proposal distribution:
The variance can have an impact. The pattern showed for the chain given with $LN\ (X_t, 1)$ as proposal distribution may show signs that are typical for the case when the variance is too high. That since the chain never converges. Perhaps a lower variance could have given better results.

## 1.4

10 MCMC sequences are generated with the generator from step 2. The respective starting points are 1,2,...,10. The Gelman-Rubin method is used for analysing the convergence of the sequences.
The output is given below and the most interesting number to interpret is the *Upper C.I.*. If this number is very close to 1, approximately around 1-1.2, convergence is concluded to have been reached. As the output shows is the *Upper C.I.* for these sequences equal to 1, hence convergence is achieved.

```
## Potential scale reduction factors:
##
##      Point est. Upper C.I.
## [1,]          1          1
```

## 1.5

Using the sample from step 1 the estimation of the integral is given by:

$$\frac{1}{n}\sum_{t=1}^{n}(X_t)$$

So the estimate of the mean given by the sample from step 1 is 3.191.
For the sample from step 2 is the first 20 observations in the sample removed due to the burn-in period. So, the estimation of the integral is calculated in the following way:

$$\frac{1}{n-20}\sum_{t=20+1}^{n}(X_t)$$

With this sample is the integral estimated to be 5.819.

## 1.6

The probability density function for a gamma distribution is:

$$\frac{1}{\Gamma(\kappa)\theta^{\kappa}}x^{\kappa-1}e^{-\frac{x}{\theta}}$$

An comparsion of the target distribution and the gamma distribution gives that $\kappa$ is equal to 6 and that $\theta$ is equal to 1.
The expected mean for a gamma distribution is given by the follwoing equation:

$$E[X] = \kappa\theta$$

Hence, the actual value of the integral is 6. The mean for step 1 is far away from this value and the mean for step 2 is very close to the actual value. This result is expected since the chain did not converge in step 1 but did so in step 2.

The next plot is used for comparing the theoretical density for *Gamma(6,1)*, the orange area, and the histogram for the values generated in step 2, the purple area. The conclusion is that the generated values follows the target distribution relatively well.
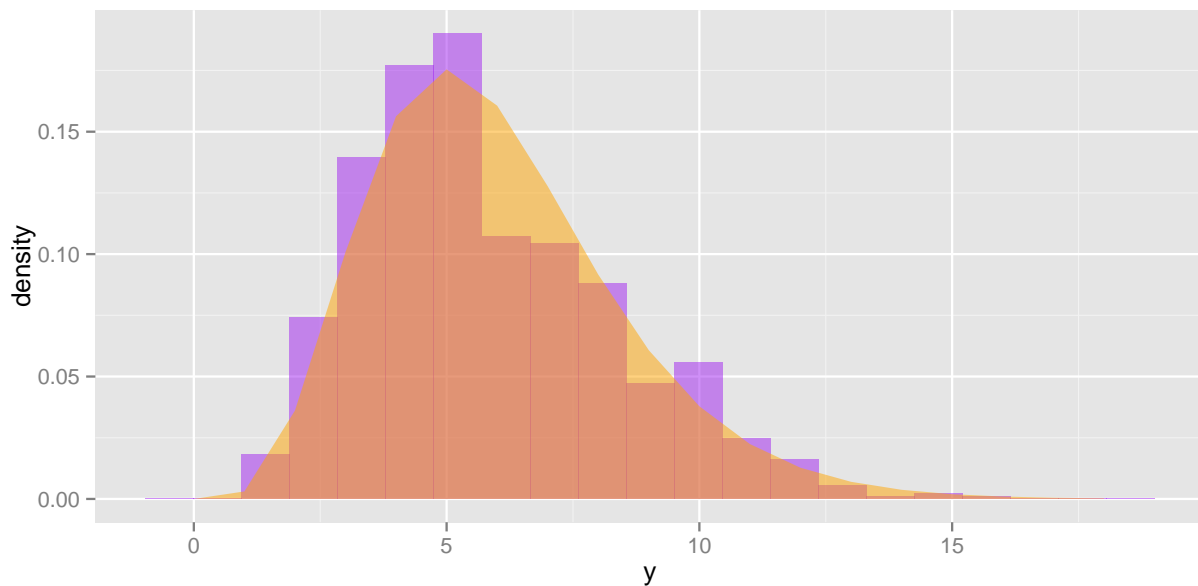


# Assignment 2

## 2.1

The dependence of the variable $y$ on the variable $x$ is illustrated with the plot below. The fits of two different models are also visualised. The blue line illustrates an quadratic model and the red line an . . . model.
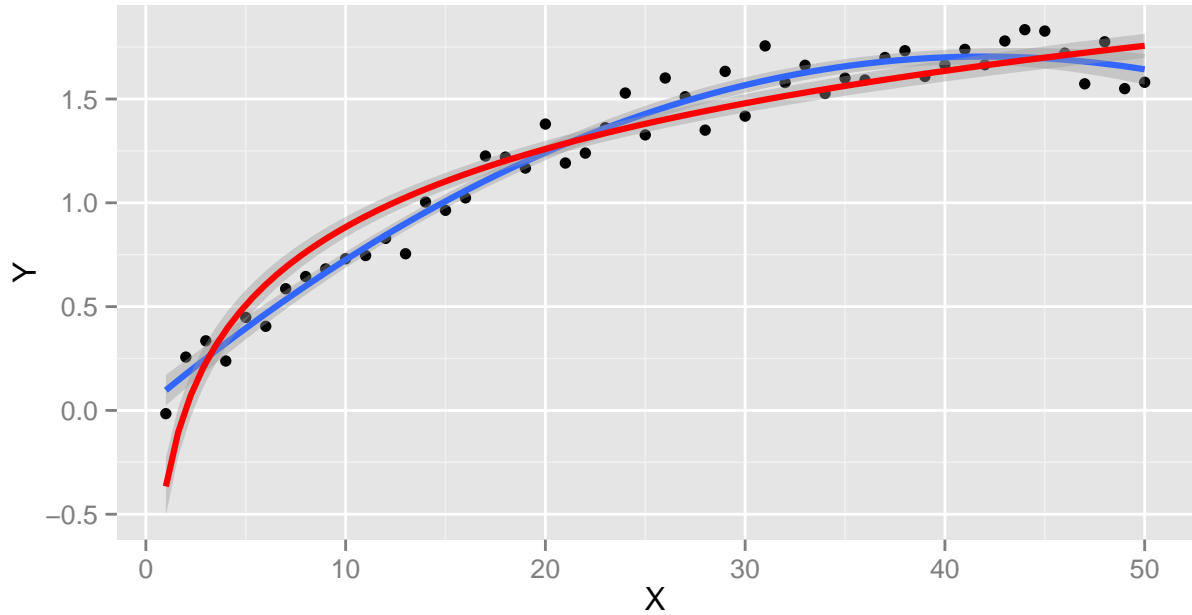
Both the quadratic model and the . . .  model seem to be relatively good fits to data. Perhaps is the quadratic model a little bit better, but that is hard to say definitely just from the plot.

## 2.2

The formula for the likelihood p($\mathbf{Y}|\mu$) is:

$$\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n exp\left(-\frac{\sum_{i=1}^{n}(Y_i - \mu_i)^2}{2\sigma^2}\right)$$

The formula for the prior $p(\mu)$ is:

$$\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{n-1} exp\left(-\frac{\sum_{i=1}^{n}(\mu_i - \mu_{i-1})^2}{2\sigma^2}\right)$$

## 2.3

The distribution for the first observation:

The distribution in the span 2-49:

$$\propto exp\left(-\frac{(\mu_i - \frac{\mu_{i-1}+Y_i+\mu_{i+1}}{3})^2}{\frac{2}{3}\sigma^2}\right)$$

The distribution for the last observation:

$$\propto exp\left(-\frac{(\mu_{50} - \frac{Y_{50}+\mu_{49}}{2})^2}{\sigma^2}\right)$$