MVE155 – Statistical inference, 14th February 2013

Computer exercise 2

Simon Sigurdhsson, 900322–0291

This is a report on exercise 8.43 in Rice (2007), which concerns a data set of interarrival times of photons. The exercise consists of 6 parts labeled (a) through (f). The exercise has been solved using the R statistical software and the code used is included throughout the report. The preamble of this code simply loads a couple of packages along with the data set:

```
library(ggplot2)
library(tikzDevice)
data <- read.csv("gamma-arrivals.txt", header=FALSE,
                 col.names=c("arrivals"))
```

Here, the `ggplot2` library is used to generate figures and the `tikzDevice` library is used to export them as Ti*k*Z code.

a. *Histogram of the interarrival times.*

```
tikz("tikz/histogram.tex", width=5, height=3)
p <- ggplot(data, aes(x=arrivals)) +
  geom_histogram(binwidth=5) + xlim(0,750)
```

As shown by figure 1 on the following page, the short interarrival times are most common, while the number of occurrences of any interarrival time decreases in an exponential manner when the interrarival time gets large. This seems to fit well with a Γ distribution with a shape parameter k close to 1.

b. *Estimated parameters for* $\Gamma(k,\theta)$*.*

```
m1 <- mean(data$arrivals)
m2 <- mean(data$arrivals^2)
moments.k <- m1^2 / (m2-m1^2)
moments.theta <- (m2 - m1^2) / m1
```

The method of moments uses the statistics $\mathrm{E}\left[X\right]$ and $\mathrm{E}\left[X^2\right]$ (with point estimates $\overline{X}$ and $\overline{X^2}$) to estimate the parameters of the distribution. For the Γ

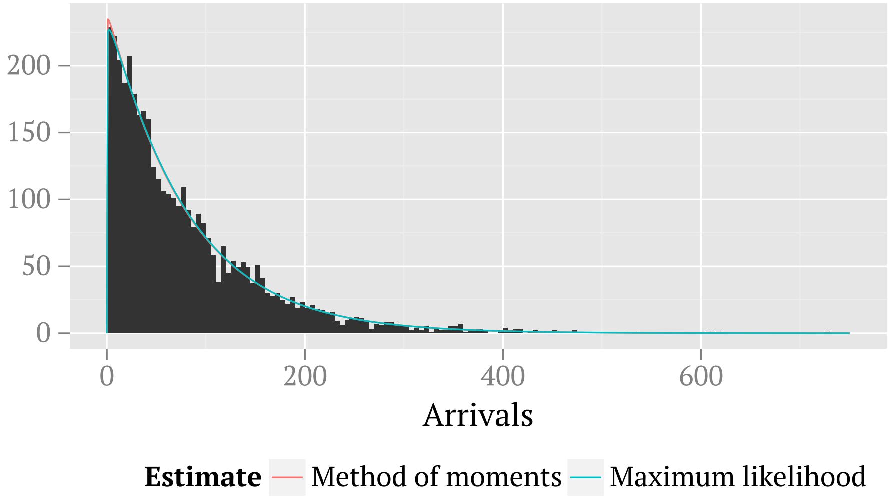

Figure 1: *Histogram of interarrival times.*

distribution, expressions for the parameters are easily derived:

$$\begin{cases} \mathrm{E}\left[X\right] = k\theta, \\ \mathrm{E}\left[X^2\right] = \theta^2 k(k+1) \end{cases} \implies \begin{cases} \hat{k} = \dfrac{\overline{X}^2}{\overline{X^2} - \overline{X}^2}, \\ \hat{\theta} = \dfrac{\overline{X^2} - \overline{X}^2}{\overline{X}} \end{cases} .$$

Using these expressions with the data set yields estimated parameter values $\hat{k}_{\mathrm{MM}} = 1.012$ and $\hat{\theta}_{\mathrm{MM}} = 78.96$, which seems to correspond to the theory that the data is Γ-distributed with k close to 1.

```
f <- function(k) abs((log(k)-digamma(k)) -
  (log(mean(data$arrivals))-mean(log(data$arrivals))))
r <- nlm(f, moments.k)
likelihood.k <- r$estimate
likelihood.theta <- mean(data$arrivals)/likelihood.k
```

The likelihood function given a Γ distribution is $L(k,\theta) = \prod_{i=1}^{N} (\Gamma(k)\theta^k)^{-1} x_i^{k-1} \exp\left(-{}^{x}\!/_{\theta}\right)$. By instead considering the log-likelihood funcion,

$$l(k,\theta) = (k-1)\sum_{i=1}^{N} \log\left(x_i\right) - \sum_{i=1}^{N} \frac{x_i}{\theta} - N k \log\left(\theta\right) - N \log\left(\Gamma(k)\right),$$

the resulting maximum likelihood problem is simplified. Taking the derivative with respect to each of the parameters k and θ and setting it to zero (thus finding the maximum) yields an estimate for θ,

$$\hat{\theta} = \frac{1}{\hat{k}N}\sum_{i=1}^{N} x_i,$$

and a formula for k,

$$\log\left(\hat{k}\right) - \frac{\Gamma'(\hat{k})}{\Gamma(\hat{k})} = \log\left(\frac{1}{N}\sum_{i=1}^{N} x_i\right) - \frac{1}{N}\sum_{i=1}^{N}\log\left(x_i\right)$$

which has no closed-form solution but from which $\hat{k}$ can be calculated numerically.

Doing so results in estimates $\hat{k}_{\mathrm{ML}} = 1.026$ and $\hat{\theta}_{\mathrm{ML}} = 77.88$,

c. *Comparing estimated distributions.*

```r
estimated.pdfs <- data.frame(x=0:750,
  moments=dgamma(0:750, moments.k, scale=moments.theta),
  likelihood=dgamma(0:750, likelihood.k, scale=likelihood.theta))
scalefactor = 5*length(data$arrivals) # 5 = binwidth
p <- p + geom_line(data=estimated.pdfs,
  aes(x=x, y=moments*scalefactor, colour='#0072B2'))
p <- p + geom_line(data=estimated.pdfs,
  aes(x=x, y=likelihood*scalefactor, colour='#D55E00'))
p + scale_colour_hue("Estimate",
  labels=c('Method of moments','Maximum likelihood')) +
  opts(legend.direction = "horizontal", legend.position = "bottom")
dev.off()
```

As shown by figure 1 on the previous page, the estimated distributions correspond well to the histogram.

d. *Estimate errors with bootstrapping.*

```r
sample.moments <- function(){
  samples <- rgamma(length(data$arrivals), moments.k, scale=moments.theta)
  m1 <- mean(samples)
  m2 <- mean(samples^2)
  k <- m1^2 / (m2-m1^2)
  theta <- (m2 - m1^2) / m1
  c(k, theta)
}
simulated.moments <- t(replicate(1000, sample.moments()))
sample.likelihood <- function(){
  samples <- rgamma(length(data$arrivals), likelihood.k, scale=likelihood.theta)
  f <- function(k) abs((log(k)-digamma(k)) -
  (log(mean(samples))-mean(log(samples))))
  r <- nlm(f, likelihood.k)
  k <- r$estimate
  theta <- mean(samples)/k
  c(k, theta)
}
simulated.likelihood <- t(replicate(1000, sample.likelihood()))
moments.k.error <- sd(simulated.moments[,1])
moments.theta.error <- sd(simulated.moments[,2])
likelihood.k.error <- sd(simulated.likelihood[,1])
likelihood.theta.error <- sd(simulated.likelihood[,2])
```

Bootstrapping the parameters by generating 1000 samples of the same size as the initial dataset reveals the error of the four estimated parameters to be

$$\begin{aligned} s_{\hat{k}_{\mathrm{MM}}} &= 0.032, & s_{\hat{\theta}_{\mathrm{MM}}} &= 2.83, \\ s_{\hat{k}_{\mathrm{ML}}} &= 0.020, & s_{\hat{\theta}_{\mathrm{ML}}} &= 1.98. \end{aligned}$$

As expected, the maximum likelihood estimate has a slightly smaller error.

e. *Confidence intervals with bootstrapping.*

```
cs <- unname(quantile(simulated.moments[,1], c(0.025, 0.975)))
moments.k.ci <- c(2*moments.k - cs[2], 2*moments.k - cs[1])
cs <- unname(quantile(simulated.moments[,2], c(0.025, 0.975)))
moments.theta.ci <- c(2*moments.theta - cs[2], 2*moments.theta - cs[1])
cs <- unname(quantile(simulated.likelihood[,1], c(0.025, 0.975)))
likelihood.k.ci <- c(2*likelihood.k - cs[2], 2*likelihood.k - cs[1])
cs <- unname(quantile(simulated.likelihood[,2], c(0.025, 0.975)))
likelihood.theta.ci <- c(2*likelihood.theta - cs[2], 2*likelihood.theta - cs[1])
```

Using the bootstrapping from (d), one can calculate confidence intervals for the parameters. What one finds is that

$$\begin{aligned} k_{\mathrm{MM}} &\in [0.948, 1.076], & \theta_{\mathrm{MM}} &\in [73.30, 84.39], \\ k_{\mathrm{ML}} &\in [0.983, 1.062], & \theta_{\mathrm{ML}} &\in [74.22, 81.67]. \end{aligned}$$

Again, quite expectedly, the maximum likelihood estimate has a tighter confidence interval, which of course is due to the smaller error.

f. *Are arrival times a Poisson process?* Since all operations above indicate that k is close to 1, let us assume that we have $k = 1$. Since it is known (and easily provable) that $\Gamma(1,\theta) \sim \mathrm{Exp}\,(1/\theta)$, and since the exponential distribution in fact describes the time between events in a Poisson process (this can be derived from the definition of a Poisson process), one can conclude that the interarrival time distribution is fairly consistent with a Poisson process model for arrival times.

References

Rice, John (2007). *Mathematical statistics and data analysis*. Australia Belmont, CA: Thompson/Brooks/Cole. ISBN: 9780495118688.