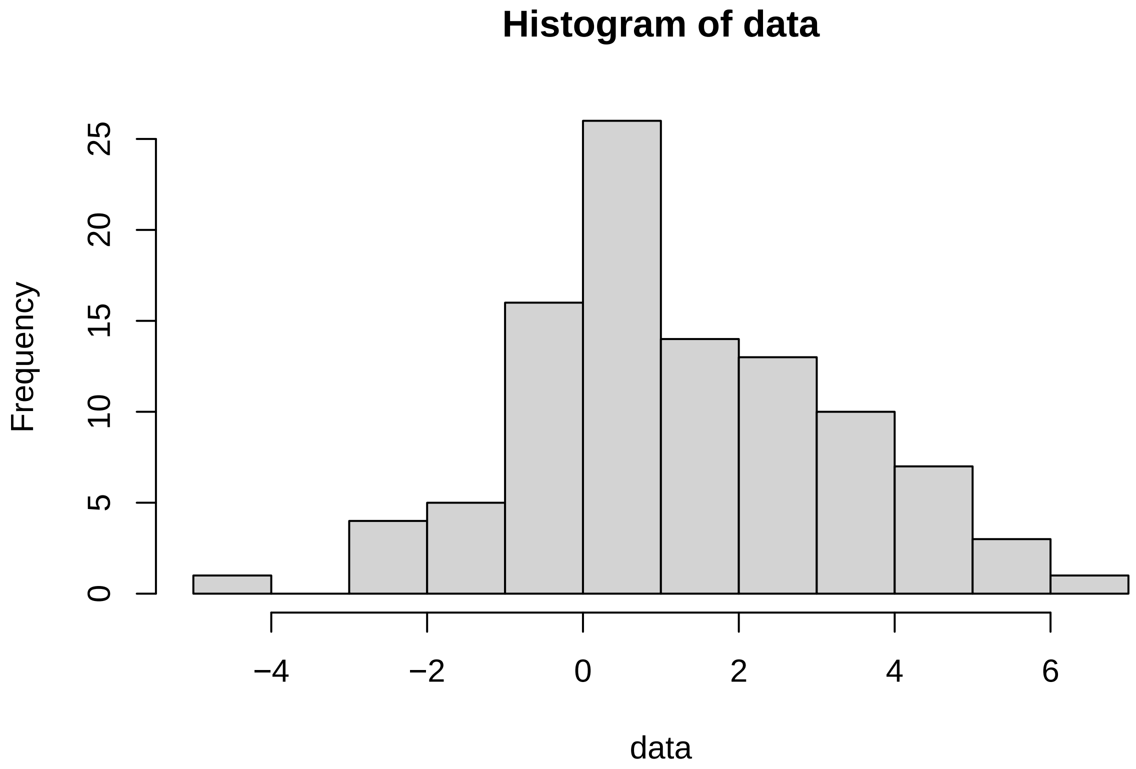
lab2 Problem2

Jaskirat S Marar

11/12/2021

Log Likelihoood estimators μ & σ

The given data is plotted for observation.



As we can see that the data seems to be normally distributed with its mean lying somewhere close to 1, which confirms what is given to us in the problem statement about the data being distributed normally. Let us proceed further with solving this problem.

Given our data, we can write the likelihood as the joint density of the sample i.e.

$$L(\mu,\sigma^2) = \prod_{i=1}^{n} f(y_i|\mu,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{y_i-\mu}{2\sigma^2}\Big) = \Big(\frac{1}{2\pi\sigma^2}\Big)^{n/2} \exp\Big(\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\Big)$$

Taking logs on both sides

$$ln[L(\mu,\sigma^2)] = -\frac{n}{2}ln\sigma^2 - \frac{n}{2}ln2\pi - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2$$

For MLE of μ & σ, we take the partial derivatives w.r.t. μ & σ and set them to zero.

First, derivating w.r.t μ, the first and second term become 0, and we derivate the 3rd term leading to:

$$\frac{\partial(ln[L(\mu,\sigma^2)])}{\partial\mu} = \frac{1}{\sigma^2}\Big(\sum_{i=1}^{n}(y_i) - n\mu\Big)$$

If we equate this to 0, and isolate μ, we get:

$$\mu = \frac{1}{n}\sum_{i=1}^{n}y_i$$

This we know to be the sample mean, we can summarize the result as follows:

$$\hat{\mu}_{MLE} = \bar{y}$$

Similarly for σ, now we derivate the log likelihood function w.r.t. σ, the second term becomes zero and we get:

$$\frac{\partial(ln[L(\mu,\sigma^2)])}{\partial\sigma} = -\frac{n}{2}.\frac{2}{\sigma} - \Big(\frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2}\Big).\Big(-\frac{2}{\sigma^3}\Big)$$

Simplifying:

$$\frac{\partial(ln[L(\mu,\sigma^2)])}{\partial\sigma} = -\frac{n}{\sigma} + \frac{\sum_{i=1}^{n}(y_i-\mu)^2}{\sigma^3}$$

Setting partial derivative to 0 and solving for σ, we get

$$\sigma^2 = \frac{\sum_{i=1}^{n}(y_i-\mu)^2}{n}$$

And this expression we know is the sample variance, so we can summarize as:

$$\hat{\sigma}^2_{MLE} = s^2$$

In terms of sample values, the variance can be written as follows:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \frac{1}{n}\sum_{i=1}^{n}y_i)^2}{n}$$

Now that we have simplified the functions for calculating the estimators, we can use the built in R functions for mean and variance for our given data:

```
mu_data = mean(data)
sigma_sq_data = var(data)

cat("mu = ", mu_data, "\n")
```

```
## mu =  1.275528
```

```
cat("sigma = ", sqrt(sigma_sq_data), "\n")
```

```
## sigma =  2.016082
```

Writing the loglikelihood function

We now write the function for calculating the log-likelihood which will be then used to optimize with initial parameters, θ

```
loglikelihood <- function(theta, x) {
  mu = theta[1]
  sig = theta[2]
  n = length(x)
  lglklhood = -(n/2)*(log(2*pi*sig^2)) + (-1/(2*sig^2))*sum((x-mu)^2)
  return(-lglklhood)
}
```

We are passing the negative value of the log-likelihood as the optim() minimizes that are passed into it. Hence, in order to maximize log-likelihood, we return the negative value of the log-likelihood.

Optimizing without specifying gradient

Next we pass the initial parameters $\mu = 0$ and $\sigma = 1$ and optimize as follows:

```
optim_norm_BFGS_withoutgradient <- optim(par = c(0,1), loglikelihood, x = data, method = "BFGS")
optim_norm_BFGS_withoutgradient
```

```
## $par
## [1] 1.275528 2.005977
##
## $value
## [1] 211.5069
##
## $counts
## function gradient
##       37       15
##
## $convergence
## [1] 0
##
## $message
## NULL
```

```
optim_norm_CG_withoutgradient <- optim(par = c(0,1), loglikelihood, x = data, method = "CG")
optim_norm_CG_withoutgradient
```

```
## $par
## [1] 1.275528 2.005977
##
## $value
## [1] 211.5069
##
## $counts
## function gradient
##      297       45
##
## $convergence
## [1] 0
##
## $message
## NULL
```

The results show that when using the BFGS method, the function was called 37 times and the calculated gradient was 15. In the case of the CG method, the function was called 297 times with the gradient being 45. The results are pretty close to the true values that were calculated earlier in the paper.

Optimizing with gradient

Now we'll attempt to optimize by providing a gradient function. For writing the gradient function we will use the understanding that we are optimizing the log-likelihood function. And we will obtain the optimized value of the log-likehood function at the MLE values for our parameters i.e. μ and σ. Thus, our gradient function will be written to return the following expressions:

$$\frac{\partial(ln[L(\mu,\sigma^2)])}{\partial\mu} = \frac{1}{\sigma^2}\Big(\sum_{i=1}^{n}(y_i) - n\mu\Big)$$

and,

$$\frac{\partial(ln[L(\mu,\sigma^2)])}{\partial\sigma} = -\frac{n}{\sigma} + \frac{\sum_{i=1}^{n}(y_i-\mu)^2}{\sigma^3}$$

So to reflect this, our code is as follows:

```
gradient <- function(par, x) {
  mu = par[1]
  sigma = par[2]
  n = length(x)
  mu_est = -(sum(x) - n*mu)/sigma^2
  sigma_est = n/sigma - sum((x - mu)^2)/sigma^3
  return(c(mu_est, sigma_est))
}
```

Note that we returned the -ve values of μ and σ. This is because we are minimizing -ve log likelihood, thus we accordingly need to pass the -ve values of estimators that will minimize the function. And thus, we modify our optim function with gradient for both methods as follows:

```
optim_norm_BFGS_withgradient <- optim(par = c(0,1),
                                      gr = gradient,
                                      loglikelihood,
                                      x = data,
                                      method = "BFGS")
optim_norm_BFGS_withgradient
```

```
## $par
## [1] 1.275528 2.005977
##
## $value
## [1] 211.5069
##
## $counts
## function gradient
##       39       15
##
## $convergence
## [1] 0
##
## $message
## NULL
```

```r
optim_norm_CG_withgradient <- optim(par = c(0,1),
                                    gr = gradient,
                                    loglikelihood,
                                    x = data,
                                    method = "CG")
optim_norm_CG_withgradient
```

```
## $par
## [1] 1.275528 2.005976
##
## $value
## [1] 211.5069
##
## $counts
## function gradient
##       53       17
##
## $convergence
## [1] 0
##
## $message
## NULL
```

The results are still accurate. All 4 of our attempts lead to convergence, with the value of the function also resulting in the same output for all 4 calls. The bigger difference actually lies in the number of times the function is getting called. If you compare the counts of the 4 calls, we see that there are significantly lesser number of calls to the function being optimized i.e. the loglikelihood() in the CG method, where the number of loglikelihood() calls have been reduced by ~80%! This obviously comes at the cost of accuracy. As we can see that even though the optimized value of mean and the resulting function values are equal to the true value, there is a slight difference in accuracy of the sigma values. And this delta was slightly higher in the CG method with gradient.

Recommended configuration?

Therefore on the matter of desired settings, because this data set is much smaller, the speed of the optimization doesn't get impacted much, but with larger multivariate datasets, an increased number of function calls can create a processing bottleneck without any significant improvement in accuracy of our results. Hence the recommendation should be that if it is convenient to analytically derive the gradient of the function to be optimized, then we can save on processing speed and get results faster without much loss in precision.And even when the gradient function is not easy, to solve there are packages available that can be used to build a gradient function.

Log-likelihood vs Likelihood?

The choice between maximizing likelihood vs log(likelihood) is largely led by the following 2 reasons:

1. Analytical convenience: Log() is a monotonically increasing function hence in order to maximize likelihood, we just have to maximize loglikehood. Also, the distributive properties of logarithms help simplify most expressions from multiplication to summation, especially as in the case of Gaussian, we dont have to calculate the exponential terms by taking logarithms. From a computational perspective, summation is not as computationally expensive as multiplication.

2. Avoiding underflow: In statistics and ML we will be working with large datasets comprising of small numbers. If we keep working with likelihood functions, the resulting product terms can very quickly

lead to extremely small numbers and loss of floating point precision thus resulting in underflow. Hence, it is recommended that we use log-likelihood for our computations.

Appendix - CODE

```r
load("data.Rdata")
#hist(data)

loglikelihood <- function(theta, x) {
  mu = theta[1]
  sig = theta[2]
  n = length(x)
  lglklhood = -(n/2)*(log(2*pi*sig^2)) + (-1/(2*sig^2))*sum((x-mu)^2)
  return(-lglklhood)
}

optim_norm <- optim(par = c(0,1), loglikelihood, x = data, method = "BFGS")
optim_norm

optim_norm1 <- optim(par = c(0,1), loglikelihood, x = data, method = "CG")
optim_norm1


mean(data)
sd(data)



gradient <- function(par, x) {
  mu = par[1]
  sigma = par[2]
  n = length(x)
  mu_est = -(sum(x) - n*mu)/sigma^2
  sigma_est = n/sigma - sum((x - mu)^2)/sigma^3
  return(c(mu_est, sigma_est))
}


optim_norm2 <- optim(par = c(0,1), gr = gradient, loglikelihood, x = data, method = "BFGS")
optim_norm2

optim_norm3 <- optim(par = c(0,1), gr = gradient, loglikelihood, x = data, method = "CG")
optim_norm3
```