HOMEWORK 5 SOLUTIONS

ALEX CHIN

1. The geometric model. The method of moments estimator sets the population mean, $1/p$, equal to the sample mean, $\bar{X} = n^{-1}\sum_{i=1}^n X_i$. Inverting to solve for p gives

$$\hat{p}_{\text{MOM}} = \bar{X}^{-1}.$$

For the maximum likelihood estimator, the likelihood is

$$L(p|X_1,\dots,X_n) = \prod_{i=1}^n (p(1-p)^{X_i-1}) = p^n(1-p)^{n(\bar{X}-1)}$$

and the log-likelihood is therefore

$$\ell(p) = n\log p + n(\bar{X}-1)\log(1-p).$$

The derivative is

$$\ell'(p) = \frac{n}{p} - \frac{n(\bar{X}-1)}{1-p}.$$

Setting equal to zero and solving for p gives

$$\hat{p}_{\text{MLE}} = \bar{X}^{-1}.$$

(We must also check that $\ell(p)$ achieves a maximum at $\bar{X}^{-1}$; this may be verified by checking that $\ell'(p)$ takes positive for $p < \bar{X}^{-1}$ and negative values for $p > \bar{X}^{-1}$.)

We can get the asymptotic distribution using the delta method. We have from the central limit theorem that

$$\sqrt{n}(\bar{X} - 1/p) \Rightarrow \mathcal{N}\left(0, \frac{1-p}{p^2}\right).$$

Taking $g(\theta) = 1/\theta$ gives $(g'(\theta))^2 = \theta^{-4}$, which for $\theta = 1/p$ is $(g'(\theta))^2 = p^4$. Hence

$$\sqrt{n}(\hat{p}_{\text{MLE}} - p) = \sqrt{n}(1/\bar{X} - p) = \sqrt{n}(g(\bar{X}) - g(1/p)) \Rightarrow \mathcal{N}\left(0, p^2(1-p)\right).$$

Alternatively, we could obtain the variance using the Fisher information:

$$\sqrt{n}(\hat{p}_{\text{MLE}} - p) \Rightarrow \mathcal{N}\left(0, \frac{1}{I(p)}\right),$$

Stats 200: Autumn 2016.

where $I(p)$ is the Fisher information for a single observation. We compute

$$\begin{aligned} I(p) = -\mathbf{E}_p[\ell''(p)] &= -\mathbf{E}_p\left[\frac{\partial^2}{\partial^2 p}\left(\log p + (X-1)\log(1-p)\right)\right] \\ &= -\mathbf{E}_p\left[\frac{\partial}{\partial p}\left(\frac{1}{p} - \frac{X-1}{1-p}\right)\right] \\ &= -\mathbf{E}_p\left[-\frac{1}{p^2} - \frac{X-1}{(1-p)^2}\right] \\ &= \frac{1}{p^2(1-p)}. \end{aligned}$$

So

$$\sqrt{n}(\hat{p}_{\text{MLE}} - p) \Rightarrow \mathcal{N}\left(0, p^2(1-p)\right).$$

2. Fisher information in the normal model.

(a) Denote $v = \sigma^2$. Then

$$f(X|\mu, v) = \frac{1}{\sqrt{2\pi v}} e^{-\frac{1}{2v}(X-\mu)^2}$$

and

$$\ell(\mu, v) = \log f(X|\mu, v) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log v - \frac{1}{2v}(X-\mu)^2.$$

In order to obtain the Fisher information matrix $I(\mu, v)$, we must compute the four second-order partial derivatives of $\ell(\mu, v)$. These quantities are

$$\begin{aligned} \frac{\partial^2 \ell}{\partial \mu^2} &= -\frac{1}{v}, \\ \frac{\partial^2 \ell}{\partial v^2} &= \frac{1}{2v^2} - \frac{1}{v^3}(X-\mu)^2, \\ \frac{\partial^2 \ell}{\partial \mu \partial v} = \frac{\partial^2 \ell}{\partial v \partial \mu} &= -\frac{X-\mu}{v^2}. \end{aligned}$$

Then

$$I(\mu, v) = -\mathbf{E}_{\mu,v}\begin{bmatrix} \frac{\partial^2 \ell}{\partial \mu^2} & \frac{\partial^2 \ell}{\partial \mu \partial v} \\ \frac{\partial^2 \ell}{\partial v \partial \mu} & \frac{\partial^2 \ell}{\partial v^2} \end{bmatrix} = \begin{bmatrix} 1/v & 0 \\ 0 & 1/2v^2 \end{bmatrix}.$$

This matrix has inverse

$$I(\mu, v)^{-1} = 2v^3 \begin{bmatrix} 1/2v^2 & 0 \\ 0 & 1/v \end{bmatrix} = \begin{bmatrix} v & 0 \\ 0 & 2v^2 \end{bmatrix}.$$

Substituting back $v = \sigma^2$, we have

$$I(\mu, \sigma^2)^{-1} = \begin{bmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{bmatrix},$$

which we conclude is the asymptotic variance of the maximum likelihood estimate. In other words,

$$\sqrt{n}\left(\begin{bmatrix}\bar{X}\\ S^2\end{bmatrix}-\begin{bmatrix}\mu\\ \sigma^2\end{bmatrix}\right)\Rightarrow\mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}\sigma^2 & 0\\ 0 & 2\sigma^4\end{bmatrix}\right).$$

(b) The joint log-likelihood in this one-parameter sub-model is given by

$$\ell(v) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log v - \frac{1}{2v}\sum_{i=1}^n X_i^2,$$

where again $v = \sigma^2$. Then

$$\ell'(v) = -\frac{n}{2v} + \frac{1}{2v^2}\sum_{i=1}^n X_i^2,$$

and setting equal to zero and solving for v gives

$$\tilde{v} = \tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n X_i^2.$$

Since the off-diagonals of the inverse Fisher information matrix are zero, the sample mean and standard deviation are asymptotically uncorrelated, and so $\hat{\sigma}^2$ and $\tilde{\sigma}^2$ have the same asymptotic standard error.

3. Necessity of regularity conditions.

(a) The likelihood is

$$L(\theta) = \frac{1}{\theta^n}\prod_{i=1}^n \mathbb{1}\{0 \le X_i \le \theta\}.$$

Now, notice that the expression

$$\prod_{i=1}^n \mathbb{1}\{0 \le X_i \le \theta\},$$

taken as a function of θ, is the same as

$$\mathbb{1}\{\theta > X_i \text{ for all } i\} = \mathbb{1}\{\theta \ge \max_i X_i\}.$$

This means that the likelihood can be written as

$$L(\theta) = \frac{1}{\theta^n}\mathbb{1}\{\theta \ge \max_i X_i\},$$

and that the maximum likelihood estimate is the value of θ that maximizes $1/\theta^n$ on the interval $[\max_i X_i, \infty)$. Since $1/\theta^n$ is a decreasing function, this maximum occurs at the left endpoint, so

$$\hat{\theta} = \max_i X_i.$$

(b) The true parameter θ must satisfy $\theta \geq X_i$ for all $i = 1, \ldots, n$, since the range of X_i is bounded above by θ. Hence $\theta \geq \max_i X_i = \hat{\theta}$ as well. This means that for any value of n, $\sqrt{n}(\hat{\theta} - \theta)$ takes on positive values with probability zero, so $\sqrt{n}(\hat{\theta} - \theta)$ cannot be asymptotically normally distributed.

4. Generalized method-of-moments and the MLE.

(a) A Poisson random variable has mass function

$$f(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{1}{x!} e^{x \log \lambda - \lambda}.$$

for $x = 0, 1, 2, \ldots$. Reparametrizing by $\theta = \log \lambda$, we obtain

$$f(x|\theta) = \frac{1}{x!} e^{\theta x - e^{\theta}},$$

which is of the form in Equation (1). The functions are given by

$$T(x) = x, \qquad A(\theta) = e^{\theta}, \qquad \text{and} \qquad h(x) = \frac{1}{x!}.$$

(b) The derivative of the right hand side is

$$\begin{aligned} \frac{d}{d\theta} \int e^{\theta T(x) - A(\theta)} h(x) dx &= \int \frac{d}{d\theta} e^{\theta T(x) - A(\theta)} h(x) dx \\ &= \int (T(x) - A'(\theta)) e^{\theta T(x) - A(\theta)} h(x) dx. \end{aligned}$$

Since the derivative of the left hand side is 0, we have

$$0 = \int (T(x) - A'(\theta)) e^{\theta T(x) - A(\theta)} h(x) dx,$$

which implies

$$\underbrace{\int T(x) e^{\theta T(x) - A(\theta)} h(x) dx}_{\mathbf{E}_\theta[T(X)]} = A'(\theta) \underbrace{\int e^{\theta T(x) - A(\theta)} h(x) dx}_{1}.$$

Using the identities noted above, we obtain the formula

$$\mathbf{E}_\theta[T(X)] = A'(\theta).$$

(By replacing integrals with sums, the identity holds for discrete models as well.)

In the Poisson model, $A(\theta) = e^{\theta}$, so $A'(\theta) = e^{\theta}$ as well, and $T(X) = X$. This means

$$\mathbf{E}_\theta[X] = e^{\theta} = \lambda,$$

which we know to be true.

(c) From part (b), $\mathbf{E}_\theta[T(X)] = A'(\theta)$, so the generalized method-of-moments estimator is the value of θ satisfying

$$A'(\theta) = \frac{1}{n}\sum_{i=1}^{n} T(X_i).$$

We now compute the maximum likelihood estimator. The log-likelihood is

$$\theta \sum_{i=1}^{n} T(X_i) - nA(\theta) + \sum_{i=1}^{n} \log h(X_i),$$

which has derivative

$$\sum_{i=1}^{n} T(X_i) - nA'(\theta).$$

Setting equal to zero, we see that the MLE must satisfy

$$A'(\theta) = \frac{1}{n}\sum_{i=1}^{n} T(X_i),$$

which is the same as the GMM estimator for $g(x) = T(x)$.

(d) In the Poisson model $T(x) = x$, so the MLE is equal to the parameter value such that $\hat{\lambda} = e^{\hat{\theta}} = \frac{1}{n}\sum_{i=1}^{n} X_i$, which defines the usual method of moments estimator.

5. Computing the gamma MLE.

(a) As in Lecture 13, we denote the function

$$f(\alpha) = \log\alpha - \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} = \log\alpha - \psi(\alpha),$$

where $\psi(\alpha) = \Gamma'(\alpha)/\Gamma(\alpha)$ is the digamma function. Its derivative is

$$f'(\alpha) = \frac{1}{\alpha} - \psi'(\alpha),$$

where $\psi'(\alpha)$ is the trigamma function.

The Newton-Raphson update rule is

$$\alpha^{(t+1)} = \alpha^{(t)} + \frac{-f(\alpha^{(t)}) + \log\bar{X} - \frac{1}{n}\sum_{i=1}^{n}\log X_i}{f'(\alpha^{(t)})}.$$

We can implement this as follows:

```
gamma.MLE = function(X) {
  ahat = compute.ahat(X)
  bhat = ahat / mean(X)

  return(c(ahat, bhat))
}

# estimate ahat by Newton-Raphson
compute.ahat = function(X) {
  a.prev = -Inf
  a = mean(X)^2 / var(X) # initial guess

  # while not converged, do Newton-Raphson update
  while(abs(a - a.prev) > 1e-12) {
    a.prev = a
    numerator = -f(a.prev) + log(mean(X)) - mean(log(X))
    denominator = f.prime(a.prev)
    a = a.prev + numerator / denominator
  }

  return(a)
}

# define some helper functions
f = function(alpha) {
  return(log(alpha) - digamma(alpha))
}

f.prime = function(alpha) {
  return(1 / alpha - trigamma(alpha))
}
```

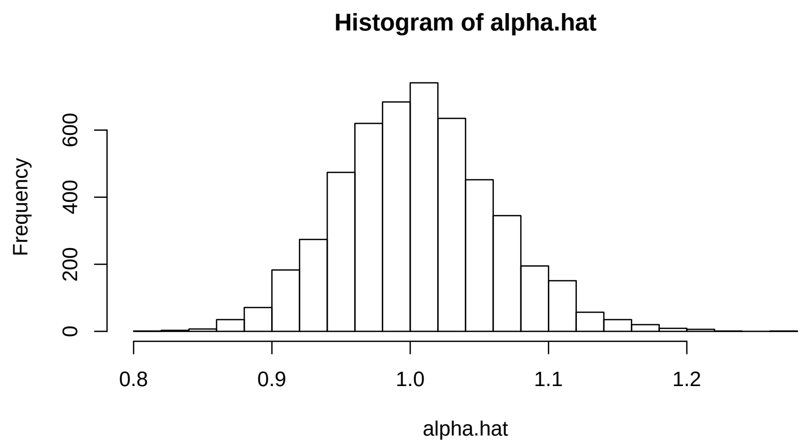
(b) We run the simulation:

```
n = 500
n.reps = 5000
alpha = 1
beta = 2

alpha.hat = numeric(n.reps)
beta.hat = numeric(n.reps)

for (i in 1:n.reps) {
  X = rgamma(n, shape = alpha, rate = beta)
  estimates = gamma.MLE(X)
```

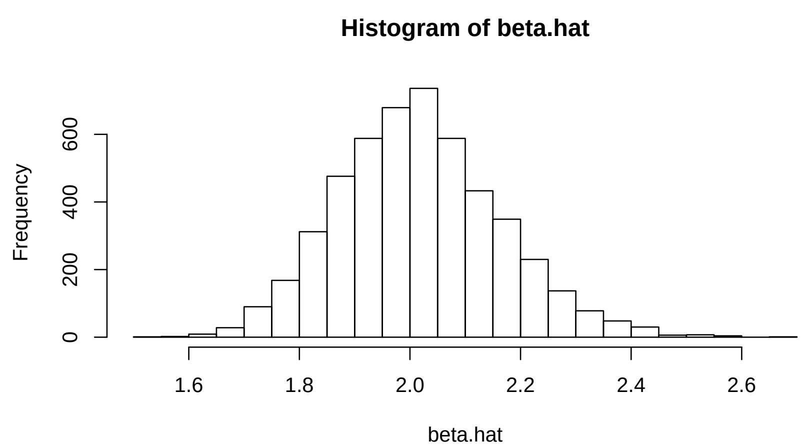
```
  alpha.hat[i] = estimates[1]
  beta.hat[i] = estimates[2]
}
```

Here are the resulting histograms:

```
hist(alpha.hat, breaks=20)
```




```
hist(beta.hat, breaks=20)
```



In class, the inverse of the Fisher information matrix was computed to be

$$I(\alpha,\beta)^{-1} = \frac{1}{\psi'(\alpha)\frac{\alpha}{\beta^2} - \frac{1}{\beta^2}} \begin{pmatrix} \frac{\alpha}{\beta^2} & \frac{1}{\beta} \\ \frac{1}{\beta} & \psi'(\alpha) \end{pmatrix}.$$

Plugging in $\alpha = 1, \beta = 2$ gives

$$I(1,2)^{-1} = \frac{4}{\psi'(1) - 1} \begin{pmatrix} 1/4 & 1/2 \\ 1/2 & \psi'(1) \end{pmatrix} \approx \begin{pmatrix} 1.551 & 3.101 \\ 3.101 & 10.202 \end{pmatrix}$$

using $\psi'(1) = \pi^2/6 \approx 1.645$.

Now, let's take a look at the empirical moments produced by the simulation. Here are the means:

```
mean(alpha.hat) # should be close to 1

## [1] 1.005632

mean(beta.hat) # should be close to 2

## [1] 2.016265
```

They are close to the true values of $\alpha = 1$ and $\beta = 2$, as expected. Now here are the variance and covariance terms:

```
var(alpha.hat)

## [1] 0.003208509

var(beta.hat)

## [1] 0.02137089

cov(alpha.hat, beta.hat)

## [1] 0.006502189
```

In order to compare to the Fisher information, we need to scale by n.

```
n * var(alpha.hat) # should be close to 1.551

## [1] 1.604255

n * var(beta.hat) # should be close to 10.202

## [1] 10.68545

n * cov(alpha.hat, beta.hat) # should be close to 3.101

## [1] 3.251095
```