HOMEWORK 7 SOLUTIONS

ALEX CHIN

1. The Laplace distribution.

(a) The joint log-likelihood is

$$\ell(\mu, b) = -n\log(2b) - \frac{1}{b}\sum_{i=1}^{n}|X_i - \mu|.$$

The likelihood is differentiable in b, so differentiating with respect to b gives

$$\frac{\partial \ell}{\partial b} = -\frac{n}{b} + \frac{1}{b^2}\sum_{i=1}^{n}|X_i - \mu|.$$

Setting this equal to 0, substituting in the MLE $\hat{\mu}$ for μ, and solving gives the MLE for b as

$$\hat{b} = \frac{1}{n}\sum_{i=1}^{n}|X_i - \hat{\mu}|.$$

We can see that the MLE $\hat{\mu}$ is the value of μ that minimizes the total absolute deviations $K(\mu) = \sum_{i=1}^{n}|X_i - \mu|$. Without loss of generality assume that the $X_1, \ldots, X_n$ are ordered. We shall see that the minimizer is the sample median $\hat{\mu} = X_m$, where $m = (n+1)/2$. When the derivative of K exists, which is everywhere except for the data points $X_1, \ldots, X_n$, it is equal to $-\sum_{i=1}^{n}\operatorname{sgn}(X_i - \mu)$, and since n is odd, this is never equal to zero. So the minimizer must occur at one of the points where the function is non-differentiable, $X_1, \ldots, X_n$. We see that $K(\mu)$ is continuous everywhere (it is the sum of absolute value functions) and furthermore it is decreasing for $\mu < X_m$ and increasing for $\mu > X_m$. Therefore the minimizer is given by $\hat{\mu} = X_m$.

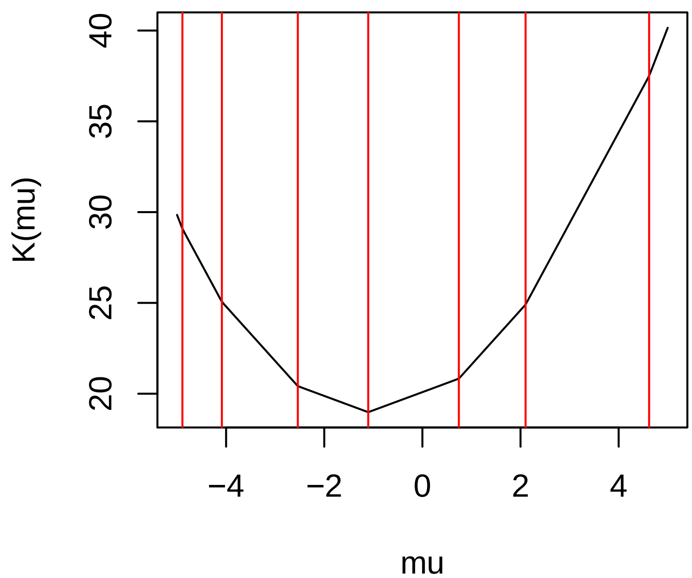
For clarity, we can plot an example showing K and K', where the red vertical lines indicate data points.

```
set.seed(13)
n = 7
x = runif(n, -5, 5)
```

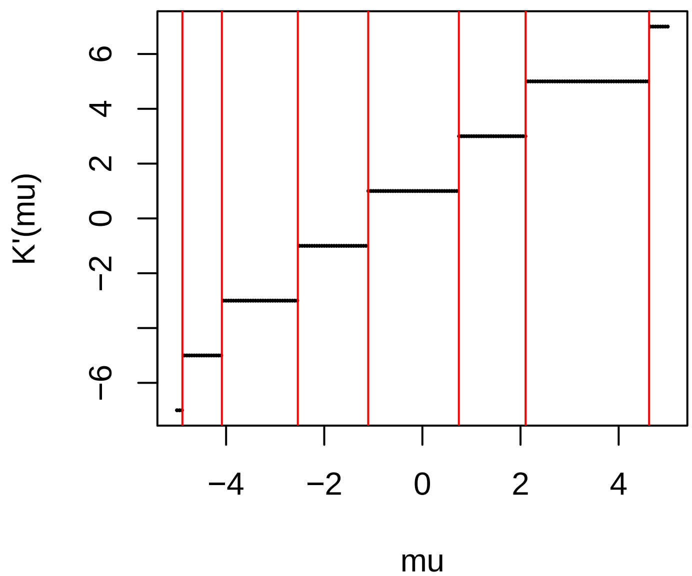
Stats 200: Autumn 2016.

```
f = function(mu) sum(abs(x - mu))
f_prime = function(mu) -sum(sign(x - mu))

mu = seq(-5, 5, 0.05)
plot(mu, sapply(mu, f), type = "l", xlab="mu", ylab="K(mu)")
for (i in x) abline(v=i, col="red")
```




```
plot(mu, sapply(mu, f_prime), xlab="mu", ylab="K'(mu)", cex=0.2, pch=20)
for (i in x) abline(v=i, col="red")
```



This estimator is more robust to outliers because it only depends on the middle few ordered values, so a few data points with extreme values won't change the median, whereas the mean depends on all data points.

(b) If $\mu = 0$ and $B \sim \text{InverseGamma}(\alpha, \beta)$, then the posterior density is given by

$$\begin{aligned} f(B|\alpha, \beta, X_1, \ldots, X_n) &\propto f(X_1, \ldots, X_n|B) f(B|\alpha, \beta) \\ &= \frac{1}{(2B)^n} \exp\left\{-\frac{1}{B}\sum_{i=1}^n |X_i|\right\} \frac{\beta^\alpha}{\Gamma(\alpha)} B^{-\alpha-1} e^{-\beta/B} \\ &\propto B^{-(\alpha+n)-1} \exp\left\{-\frac{1}{B}\left(\beta + \sum_{i=1}^n |X_i|\right)\right\}, \end{aligned}$$

where we have dropped any normalizing constants into the proportionality term. From here, we can see that the posterior distribution of B follows an InverseGamma$(\alpha + n, \beta + \sum |X_i|)$ distribution, and therefore has posterior mean $(\beta + \sum |X_i|)/(\alpha + n - 1)$.

(c) The MLE for b when $\mu = 0$ is

$$\hat{b} = \frac{1}{n}\sum_{i=1}^n |X_i|.$$

We can write the posterior mean as a weighted average

$$\underbrace{\frac{\beta + \sum_{i=1}^n |X_i|}{\alpha + n - 1}}_{\text{posterior mean}} = \frac{\alpha - 1}{\alpha + n - 1} \underbrace{\frac{\beta}{\alpha - 1}}_{\text{prior mean}} + \frac{n}{\alpha + n - 1} \underbrace{\frac{1}{n} \sum_{i=1}^n |X_i|}_{\text{MLE}}$$

of the prior mean and the MLE, from which we see that the posterior mean tends to the MLE as $n \to \infty$.

2. Bayesian inference for multinomial proportions.

(a) The posterior distribution has density proportional to

$$P_1^{\alpha_1 - 1} \times \cdots \times P_6^{\alpha_6 - 1} \times P_1^{X_1} \times \cdots \times P_6^{X_6} = P_1^{\alpha_1 + X_1 - 1} \times \cdots \times P_6^{\alpha_6 + X_6 - 1}.$$

So the posterior distribution of $(P_1, \ldots, P_6)$ given $(X_1, \ldots, X_6)$ is Dirichlet$(\alpha_1 + X_1, \ldots, \alpha_6 + X_6)$. The posterior mean and variance are given by

$$\mathbf{E}[P_i | X_1, \ldots, X_6] = \frac{\alpha_i + X_i}{\alpha_0 + n\bar{X}} \qquad \mathbf{V}[P_i | X_1, \ldots, X_6] = \frac{(\alpha_i + X_i)(\alpha_0 + n\bar{X} - \alpha_i - X_i)}{(\alpha_0 + n\bar{X})^2 (\alpha_0 + n\bar{X} + 1)}.$$

(b) We would like to select the parameters α_i such that

- The prior mean is $1/6$ for each i, and
- The prior variance is small.

Since $\mathbf{E}[P_i] = \alpha_i / \sum_{j=1}^6 \alpha_j$, a prior mean of $1/6$ can be achieved by setting $\alpha_i = \alpha$ for each i. Then the variance is given by

$$\mathbf{V}[P_i] = \frac{\alpha(6\alpha - \alpha)}{(6\alpha)^2(\alpha + 1)} = \frac{5}{36(\alpha + 1)},$$

from which we see that a large value of α achieves small variance. (The stronger our prior belief that the die is fair, the larger we would set α.)

(c) From the posterior mean calculated in part (a), we can interpret the parameters α_i as "prior counts" so an uninformative prior sets $\alpha_i = 0$. Then the posterior mean is

$$\mathbf{E}[P_i | X_1, \ldots, X_6] = \frac{X_i}{\sum_{j=1}^n X_j} = \frac{X_i}{n},$$

which is the same as the MLE (see Lecture 13).

3. GLRT and the t-test. The log-likelihood for the full model is

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

and the MLEs for μ and σ are

$$\hat{\mu} = \bar{X} \qquad \text{and} \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\hat{\mu})^2.$$

Under the submodel defined by $\mu = 0$, the log-likelihood is

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

and the MLE for σ^2 is

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

Therefore, the GLRT statistic is given by

$$\begin{aligned}
\frac{\sup_{\sigma^2}\ell(0,\sigma^2)}{\sup_{\mu,\sigma^2}\ell(\mu,\sigma^2)} &= \frac{(2\pi\tilde{\sigma}^2)^{-n/2}\exp\left\{-\frac{1}{2\tilde{\sigma}^2}\sum_{i=1}^{n}X_i^2\right\}}{(2\pi\hat{\sigma}^2)^{-n/2}\exp\left\{-\frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n}(X_i-\hat{\mu})^2\right\}} \\
&= \left(\frac{\hat{\sigma}^2}{\tilde{\sigma}^2}\right)^{n/2} \\
&= \left(\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sum_{i=1}^{n}X_i^2}\right)^{n/2} \\
&= \left(\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2 + n\bar{X}^2}\right)^{n/2}.
\end{aligned}$$

We can rewrite this as

$$\begin{aligned}
\Lambda(X_1,\ldots,X_n) &= \left(\frac{(n-1)S_X^2}{(n-1)S_X^2 + n\bar{X}^2}\right)^{n/2} \\
&= \left(\frac{n-1}{n-1+n\bar{X}^2/S_X^2}\right)^{n/2} \\
&= \left(\frac{n-1}{n-1+T^2}\right)^{n/2},
\end{aligned}$$

which is a decreasing function of T^2 and hence of $|T|$ as well.

4. Migration rates. The full log-likelihood is proportional to

$$\sum_{1 \le i,j \le 3} N_{ij} \log p_{ij}$$

and so the MLEs are given by

$$\hat{p}_{ij} = \frac{N_{ij}}{n} \qquad \text{for } 1 \le i, j \le 3$$

(see Example 13.4 in the lecture notes).

Under the equilibrium null hypothesis, the likelihood is

$$p_{11}^{N_{11}} p_{22}^{N_{22}} p_{33}^{N_{33}} p_{12}^{N_{12}+N_{21}} p_{13}^{N_{13}+N_{31}} p_{23}^{N_{23}+N_{32}} = \prod_{i=1}^{3} p_{ii}^{N_{ii}} \prod_{1 \le i < j \le 3} p_{ij}^{N_{ij}+N_{ji}}.$$

So we wish to maximize

$$\sum_{i=1}^{3} N_{ii} \log p_{ii} + \sum_{1 \le i < j \le 3} (N_{ij}+N_{ji}) \log p_{ij} + \lambda(p_{11}+p_{22}+p_{33}+2p_{12}+2p_{13}+2p_{23}-1)$$

where the last term is the Lagrange multiplier for the constraint that the parameters sum to one.

Taking derivatives and solving gives $\tilde{p}_{ii} = -N_{ii}/\lambda$ for the diagonal elements and $\tilde{p}_{ij} = -(N_{ij} + N_{ji})/(2\lambda)$ for the off-diagonal elements. We furthermore see that $\lambda = -n$ satisfies the constraint, so our MLE estimates in the submodel are

$$\tilde{p}_{ii} = \frac{N_{ii}}{n} \quad \text{for } i = j, \qquad \text{and} \qquad \tilde{p}_{ij} = \frac{N_{ij} + N_{ji}}{2n} \quad \text{for } i \ne j.$$

The generalized likelihood ratio test statistic is given by

$$\Lambda = \prod_{1 \le i,j \le n} \left(\frac{\tilde{p}_{ij}}{\hat{p}_{ij}} \right)^{N_{ij}},$$

Since $\tilde{p}_{ij} = \hat{p}_{ij}$ if $i = j$, we need only worry about the off-diagonal terms. So

$$-2 \log \Lambda = 2 \sum_{i \ne j} N_{ij} \log \frac{2N_{ij}}{N_{ij} + N_{ji}}.$$

To perform the test if n is large, we can compute $-2 \log \Lambda$ and compare it to the $(1 - \alpha)$ cutoff value of a χ^2_3 distribution, where there are 3 degrees of freedom because the submodel contains 3 fewer parameters than the full model.