Homework 5

Ze Yang (zey@andrew.cmu.edu)

Due Thursday, December 7 at 3:00 PM

You can submit separate pdf files, one generated from the R Markdown, and the other from the "derivations" required in Questions 2 and 3. The relevant .Rmd file should also be submitted.

Please do not submit photos of your homework. Scanners are available for your use.

Question 1

Redo the analysis on the Kellogg's data that we did during lecture, but this time look at the **weekly** log returns instead of the daily log returns. Show all of the R code needed to perform the analysis.

1. Reconstruct the plot comparing the five fitted models that we considered during lecture: normal, GED, nonstandard T, skewed T, and skewed T with shift.
2. Which of the five models is preferred by AIC?

```r
Kellogg =getSymbols("K", from="2010-1-1",to="2016-12-31", auto.assign=F)
```

```
## 'getSymbols' currently uses auto.assign=TRUE by default, but will
## use auto.assign=FALSE in 0.5-0. You will still be able to use
## 'loadSymbols' to automatically load data. getOption("getSymbols.env")
## and getOption("getSymbols.auto.assign") will still be checked for
## alternate defaults.
##
## This message is shown once per session and may be disabled by setting
## options("getSymbols.warning4.0"=FALSE). See ?getSymbols for details.

##
## WARNING: There have been significant changes to Yahoo Finance data.
## Please see the Warning section of '?getSymbols.yahoo' for details.
##
## This message is shown once per session and may be disabled by setting
## options("getSymbols.yahoo.warning"=FALSE).
```

```r
ldr =data.frame(dailyReturn(Ad(Kellogg), type="log"))
lwr =data.frame(weeklyReturn(Ad(Kellogg), type="log"))
```

```r
# generate MLE estimates for different distributional models
source("http://www.stat.cmu.edu/~cschafer/MSCF/ModelSelectionExample.txt")

# GED
gedout = FitGED(lwr$weekly.returns)
```

```r
names(gedout$mle) = c("mean", "sd", "nu")
print(gedout$mle)
```

```
##        mean          sd          nu
## 0.002425106 0.019569354 1.308185257
```

```r
# non-standard student.t
nonstdTout = FitNonStdT(lwr$weekly.returns)
names(nonstdTout$mle) = c("mean", "sd", "nu")
print(nonstdTout$mle)
```

```
##        mean          sd          nu
## 0.002408331 0.019835619 5.240941438
```

```r
# skewed t
skewtout = FitSkewT(lwr$weekly.returns)
print(skewtout$mle)
```

```
##            k            n       lambda       sigma2
## 2.1941746738 4.8452173405 0.0385915399 0.0003983916
```

```r
# shifted skewed t
skewtshiftout = FitSkewT(
  lwr$weekly.returns,allowshift = T, control=list(maxit=1000))
names(skewtshiftout$mle) = c("k", "n", "lambda","sigma2","shift")
print(round(skewtshiftout$mle,6))
```

```
##         k         n    lambda    sigma2     shift
##  1.907886  5.890992 -0.172947  0.000393  0.006519
```

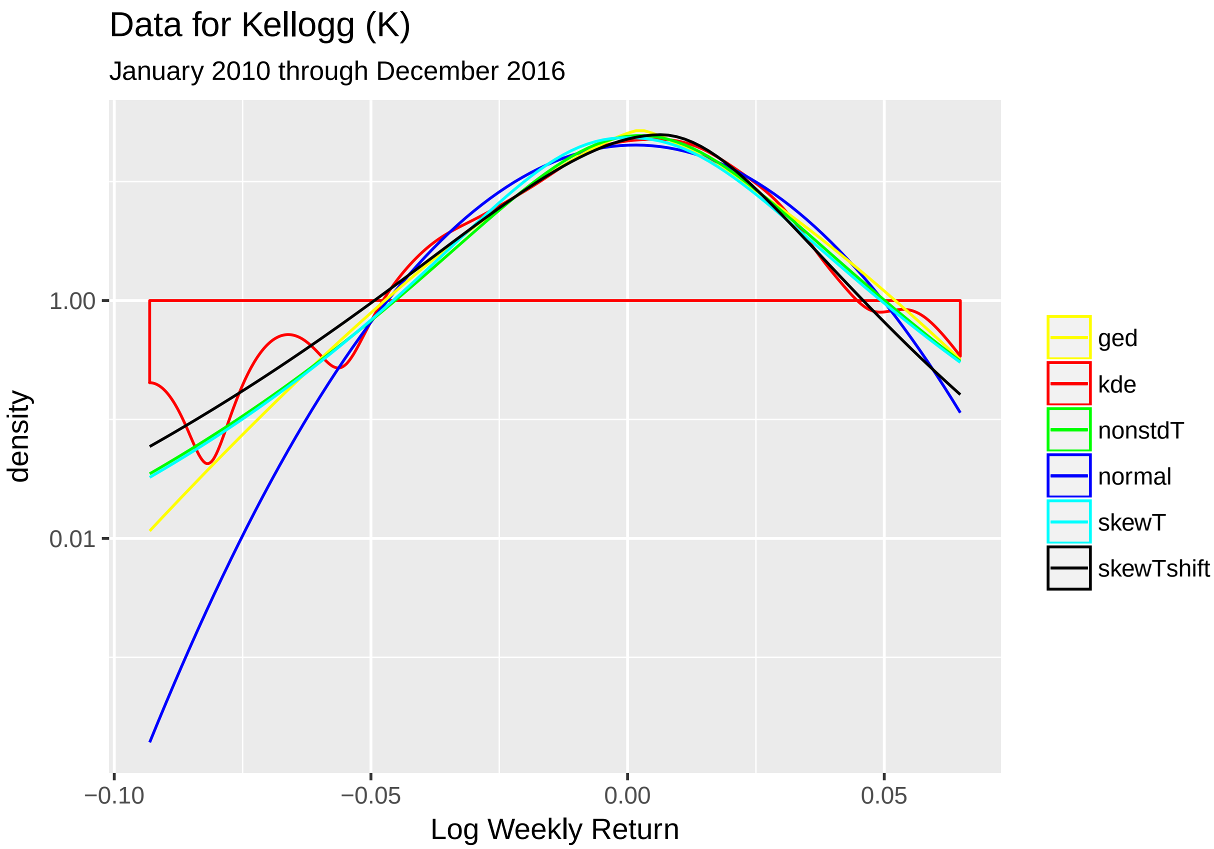
```r
# normal
n = length(lwr$weekly.returns)
normalout = c(mean(lwr$weekly.returns),
              sqrt(var(lwr$weekly.returns)*(n-1)/n))
names(normalout) = c("mean","sd")
print(normalout)
```

```
##        mean          sd
## 0.001490615 0.019670283
```

(1)

```r
# reproduce the fitted distribution plot
ggplot(lwr,aes(x=weekly.returns)) +
    geom_density(bw="SJ",aes(color="kde")) +
    stat_function(fun=dnorm, aes(color="normal"),
        args=list(
          mean=normalout[1],
          sd=normalout[2]
        )) +
  stat_function(fun=dged, aes(color="ged"),
```

```r
      args=list(
        mean=gedout$mle[1],
        sd=gedout$mle[2],
        nu=gedout$mle[3]
      )) +
stat_function(fun=dstd, aes(color="nonstdT"),
      args=list(
        mean=nonstdTout$mle[1],
        sd=nonstdTout$mle[2],
        nu=nonstdTout$mle[3]
      )) +
stat_function(fun=dSkewT, aes(color="skewT"),
      args=list(
        k=skewtout$mle[1],
        n=skewtout$mle[2],
        lambda=skewtout$mle[3],
        sigma2=skewtout$mle[4]
      )) +
stat_function(fun=dSkewT, aes(color="skewTshift"),
      args=list(
        k=skewtshiftout$mle[1],
        n=skewtshiftout$mle[2],
        lambda=skewtshiftout$mle[3],
        sigma2=skewtshiftout$mle[4],
        shift=skewtshiftout$mle[5]
      )) +
scale_color_manual(name="",
      values=c("kde"="red",
               "normal"="blue",
               "ged"="yellow",
               "nonstdT"="green",
               "skewT"="cyan",
               "skewTshift"="black")) +
  labs(x="Log Weekly Return",
       title="Data for Kellogg (K)",
       subtitle="January 2010 through December 2016") +
  scale_y_log10()
```



(2)

```r
aic = function(pdf, sample, params) {
  #' The Akaike information criterion.
  #' @param pdf: a density function for which
  #'   the likelihood is evaluated.
  #' @param sample: the observed x sample.
  #' @param params: the model parameters
  #' @return The AIC quantity.
  args = as.list(params)
  args$x = sample
  args$log = T
  log.likelihood = sum(do.call(pdf, args=args))
  return(-2*log.likelihood + 2*length(params))
}
```

```r
aic.eval = list(
  ged=aic(dged, lwr$weekly.returns, gedout$mle),
  nonstd.t=aic(dstd, lwr$weekly.returns, nonstdTout$mle),
  skew.t=aic(dSkewT, lwr$weekly.returns, skewtout$mle),
  skew.t.shift=aic(dSkewT, lwr$weekly.returns, skewtshiftout$mle),
  normal=aic(dnorm, lwr$weekly.returns, normalout)
)
```

```
aic.eval
```

```
## $ged
## [1] -1846.876
##
## $nonstd.t
## [1] -1850.368
##
## $skew.t
## [1] -1843.426
##
## $skew.t.shift
## [1] -1851.83
##
## $normal
## [1] -1828.087
```

Based on the smallest AIC, the best choice of model is the skewed t distribution with shift. The AIC value of that is -1851.83.

Question 2

Consider the example in lecture where we estimated the pair (α, σ^2) from the geometric Brownian motion. Suppose that instead I wanted to estimate (α, σ), i.e., I want to determine the asymptotic distribution of the MLE for this parameter vector.

What is the asymptotic distribution of the MLE $(\widehat{\alpha}, \widehat{\sigma})$? (Don't just say it's normal. Derive the covariance matrix.)

Hint: This is a simple exercise, don't make this more work than it needs to be. Take the result we derived in lecture, and use it as a starting point.

Let $Y_i = \log\left(\frac{S(t_{i-1}+\delta t)}{S(t_{i-1})}\right)$ be our observed log return over $[0, T]$, $\delta t = T/n$, $i = 1, 2..., n$. By the property of GBM:

$$Y_i \sim \text{i.i.d. Normal}(\mu\delta t, \sigma^2\delta t)$$

Let $\bar{Y} = \sum_{i=1}^n Y_i$ be the sample mean, $s_Y^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \bar{Y})^2$ the sample variance. \ By the results we've derived in lecture the MLE for $(\alpha, \sigma^2) = (\mu + \sigma^2/2, \sigma^2)$ is given by:

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\sigma}^2 \end{pmatrix} = \begin{pmatrix} \frac{\bar{Y}}{\delta t} + \frac{\frac{n-1}{n}s_Y^2}{2\delta t} \\ \\ \frac{\frac{n-1}{n}s_Y^2}{\delta t} \end{pmatrix} = \begin{pmatrix} \frac{n\bar{Y}}{T} + \frac{(n-1)s_Y^2}{2T} \\ \\ \frac{(n-1)s_Y^2}{T} \end{pmatrix} \tag{1}$$

And the asymptotic covariance matrix is

$$\frac{\boldsymbol{\Sigma}}{n} = \begin{pmatrix} \frac{\sigma^2}{T} + \frac{\sigma^4}{2n} & \frac{\sigma^4}{n} \\ \\ \frac{\sigma^4}{n} & \frac{2\sigma^4}{n} \end{pmatrix} \tag{2}$$

Now we are intersted in estimating (α, σ). We have

$$(\alpha, (\sigma^2)^{\frac{1}{2}}) = \boldsymbol{g}(\alpha, \sigma^2) \tag{3}$$

$$\nabla \boldsymbol{g} = \begin{pmatrix} 1 & 0 \\ \\ 0 & \frac{1}{2}(\sigma^2)^{-\frac{1}{2}} \end{pmatrix} \tag{4}$$

By the Delta method, the covariance matrix for $(\hat{\alpha}, \hat{\sigma})$ is therefore

$$\nabla \boldsymbol{g} \left(\frac{1}{n}\boldsymbol{\Sigma}\right) \nabla \boldsymbol{g}^\top = \begin{pmatrix} \frac{\sigma^2}{T} + \frac{\sigma^4}{2n} & \frac{\sigma^3}{2n} \\ \\ \frac{\sigma^3}{2n} & \frac{\sigma^2}{2n} \end{pmatrix} \tag{5}$$

And the estimator is

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\sigma} \end{pmatrix} = \begin{pmatrix} \frac{n\bar{Y}}{T} + \frac{(n-1)s_Y^2}{2T} \\ \\ \sqrt{\frac{(n-1)s_Y^2}{T}} \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix} \mu + \frac{1}{2}\sigma^2 \\ \\ \sigma \end{pmatrix}, \quad \begin{pmatrix} \frac{\sigma^2}{T} + \frac{\sigma^4}{2n} & \frac{\sigma^3}{2n} \\ \\ \frac{\sigma^3}{2n} & \frac{\sigma^2}{2n} \end{pmatrix}\right) \tag{6}$$

Question 3

Suppose that $X_1, X_2, \ldots, X_n$ are iid random variables. Each X_i takes three possible values (call these 1, 2, and 3, but the names are unimportant). The random variables are such that

$$P(X_i = 1) = p_1$$

and

$$P(X_i = 2) = p_2,$$

and, of course,

$$P(X_i = 3) = 1 - p_1 - p_2.$$

The restrictions

$$p_1 + p_2 < 1, \quad p_1 > 0, \quad p_2 > 0$$

are placed on the parameters. (This is a special case of the *multinomial distribution*, which is a generalization of the binomial distribution.)

1. Derive the MLE for $\theta = (p_1, p_2)$. (Use calculus. You do not need to verify that the Hessian is positive definite.)

2. What is the asymptotic distribution of the MLE $(\widehat{p}_1, \widehat{p}_2)$? (Don't just say it's normal. Derive the covariance matrix.)

Hint: Start by defining n_j to be the number of the X_i which are equal to j for $j = 1, 2, 3$. Of course, $n = n_1 + n_2 + n_3$. The likelihood function is then

$$L(\theta) = p_1^{n_1} p_2^{n_2} (1 - p_1 - p_2)^{n_3}.$$

Extra Practice: Derive a $100(1-\alpha)\%$ confidence interval for p_1/p_2.

(a) The log likelihood function is

$$\log \mathcal{L}(p_1, p_2|X) = n_1 \log p_1 + n_2 \log p_2 + n_3 \log(1 - p_1 - p_2) \tag{7}$$

First order condition:

$$\nabla \log \mathcal{L}(p_1, p_2|X) = \begin{pmatrix} \frac{n_1}{p_1} - \frac{n_3}{1-p_1-p_2} \\ \\ \frac{n_2}{p_2} - \frac{n_3}{1-p_1-p_2} \end{pmatrix} = \mathbf{0} \tag{8}$$

solve the system we get:

$$\begin{pmatrix} \hat{p}_1 \\ \hat{p}_2 \end{pmatrix} = \begin{pmatrix} \frac{n_1}{n} \\ \frac{n_2}{n} \end{pmatrix} = \begin{pmatrix} \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{X_i = 1\}} \\ \\ \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{X_i = 2\}} \end{pmatrix} \tag{9}$$

where $n = n_1 + n_2 + n_3$. \ **(b)** The Hessian of $\log \mathcal{L}(\boldsymbol{p}|X)$:

$$\nabla^2 \log \mathcal{L}(\boldsymbol{p}|X) = \begin{pmatrix} \frac{\partial^2 \log \mathcal{L}(\boldsymbol{p}|X)}{\partial p_1^2} & \frac{\partial^2 \log \mathcal{L}(\boldsymbol{p}|X)}{\partial p_1 \partial p_2} \\ \frac{\partial^2 \log \mathcal{L}(\boldsymbol{p}|X)}{\partial p_1 \partial p_2} & \frac{\partial^2 \log \mathcal{L}(\boldsymbol{p}|X)}{\partial p_2^2} \end{pmatrix} = \begin{pmatrix} -\frac{n_1}{p_1^2} - \frac{n_3}{(1-p_1-p_2)^2} & -\frac{n_3}{(1-p_1-p_2)^2} \\ \\ -\frac{n_3}{(1-p_1-p_2)^2} & -\frac{n_2}{p_2^2} - \frac{n_3}{(1-p_1-p_2)^2} \end{pmatrix} \tag{10}$$

Therefore

$$n\boldsymbol{I}(\boldsymbol{p}) = \mathbb{E}\left[-\nabla^2 \log \mathcal{L}(\boldsymbol{p}|X)\right] = n \begin{pmatrix} \frac{1}{p_1} + \frac{1}{1-p_1-p_2} & \frac{1}{1-p_1-p_2} \\ \\ \frac{1}{1-p_1-p_2} & \frac{1}{p_2} + \frac{1}{1-p_1-p_2} \end{pmatrix} \tag{11}$$

$$\frac{1}{n}\boldsymbol{I}^{-1}(\boldsymbol{p}) = \frac{1}{n} \begin{pmatrix} p_1(1-p_1) & -p_1 p_2 \\ \\ -p_1 p_2 & p_2(1-p_2) \end{pmatrix} \tag{12}$$

So by the property of MLE, the asymptotic distribution of $\hat{\boldsymbol{p}}$ is

$$\hat{\boldsymbol{p}} \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix} p_1 \\ p_2 \end{pmatrix}, \quad \frac{1}{n}\begin{pmatrix} p_1(1-p_1) & -p_1p_2 \\ \\ -p_1p_2 & p_2(1-p_2) \end{pmatrix}\right) \tag{13}$$

(c) Let $f(p_1, p_2) = p_1/p_2$. Then

$$\nabla f(\boldsymbol{p}) = \begin{pmatrix} \frac{1}{p_2} & -\frac{p_1}{p_2^2} \end{pmatrix} \tag{14}$$

By the Delta method, the asymptotic variance of f is

$$V = \nabla f(\boldsymbol{p}) \left(\frac{1}{n}\boldsymbol{I}^{-1}(\boldsymbol{p})\right) \nabla f(\boldsymbol{p})^\top = \frac{p_1(p_1+p_2)}{np_2^3} \tag{15}$$

Hence

$$f(\hat{\boldsymbol{p}}) = \frac{\hat{p}_1}{\hat{p}_2} \xrightarrow{d} \mathcal{N}\left(\frac{p_1}{p_2}, \frac{p_1(p_1+p_2)}{np_2^3}\right) \tag{16}$$

And of course one can replace $\boldsymbol{p}$ with $\hat{\boldsymbol{p}}$ to get an approximation of the asymptotic covariance matrix using the data, where $\hat{p}_k = n_k/n = \frac{1}{n}\sum_{i=1}^n 1_{\{X_i=k\}}$, $k = 1, 2, 3$.

The $100(1-\alpha)\%$ confidence interval is therefore (approximately, by replacing $\boldsymbol{p}$ with $\hat{\boldsymbol{p}}$)

$$CI(f(\boldsymbol{p}); \alpha) = \left[\frac{\hat{p}_1}{\hat{p}_2} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1(\hat{p}_1+\hat{p}_2)}{n\hat{p}_2^3}}, \quad \frac{\hat{p}_1}{\hat{p}_2} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1(\hat{p}_1+\hat{p}_2)}{n\hat{p}_2^3}}\right] \tag{17}$$

Where $\hat{p}_k = n_k/n = \frac{1}{n}\sum_{i=1}^n 1_{\{X_i=k\}}$, $k = 1, 2, 3$.