1. *Suppose we have an intrinsically stationary process with semivariogram:*

   for any sites si,i = 1,,n and for any constants ai,i = 1,,n with $\Sigma_{i=1,...,n}$ *ai = 0 but you did it under an assumption of second order stationarity. We will now establish it in general.*

   (a) *First show that:*

   (b) *Now take expectations of both sides to establish the result.*

2. *Let $X_o \sim Gamma(\alpha, \beta)$ with the parameterization:*

and 0 elsewhere. Let $X_i \sim \gamma(\alpha_i, \beta)$ for i = 1, , n. We construct a one-dimensional regularly spaced random field at locations i = 1, , n

$$Z(s_i) = X_o + X_i; i = i, ..., n$$

You can assume that $X_o, X_1, ..., X_n$ are independent.

(a) *What is the distirbution of $Z(s_i)$?*

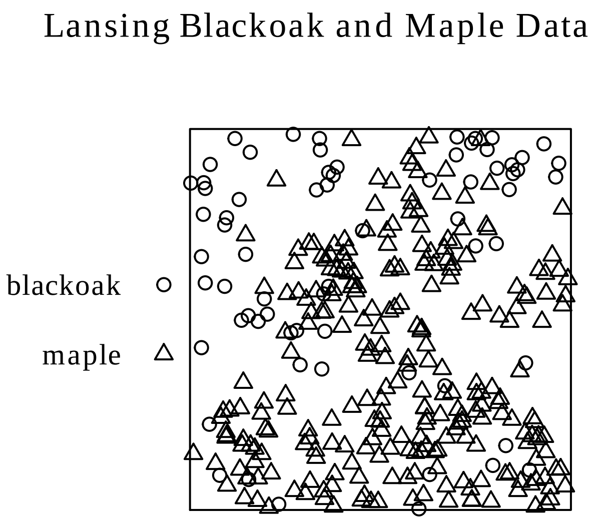
(b) *Find E(Z(si)) and Var(Z(Si)). You can use known properties of the Gamma distribution to answer this question, i.e. you can just write down the answer if you know it or can find it.*

(c) *Find $Cov[Z(S_i), Z(s_j)]$.*

(d) *Is this a second-order stationary process? Justify your answer.*

3. *The lansing data set in the spatstat package contains spatial locations of several different species of trees. We will be looking and comparing the distributions of black oaks and maples.*

```
data(lansing)
blackoak<-split(lansing)$blackoak
maple<-split(lansing)$maple

lansing.sub <- lansing[which(lansing$marks == c("blackoak", "maple"))]
lansing.sub$marks <- factor(lansing.sub$marks, levels = c("blackoak", "maple"))
plot(lansing.sub, main = "Lansing Blackoak and Maple Data")
```



The rectangular region is a unit square. Use the isotropic edge corrected version when applicable below. Answer the following questions. You will be computing several simulation envelopes below. Be patient and keep nsim=99, the default.

(a) *What does the K function measure?*

The K function measures second order properties (var/cov) of a spatial point process.

According to the help file, the K function estimates the "inter-dependence" or "clustering" of a stationary point pattern dataset. The estimate of K is used to infer the spatial pattern.

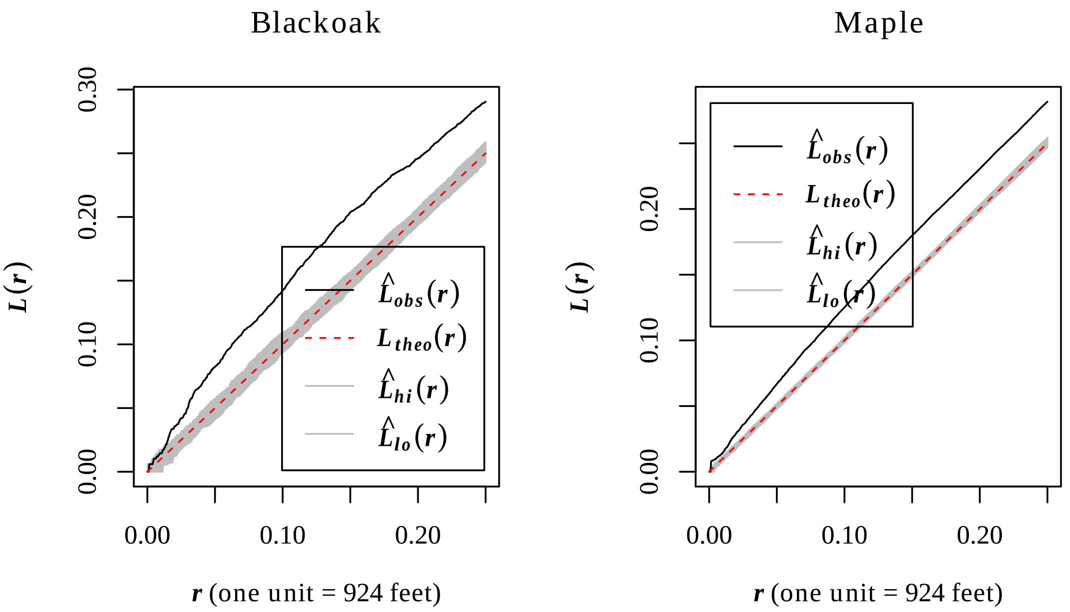
Paraphrased from the help file:

Where λ is the intensity of the process and r is the distance, λK(r) is the expected numer of additional random points within a distance of r of a typical random point of X.

(b) *It is often easier to interpret the L function than the K function. Based on the L function do the blackoaks appear to be clustered or do they appear to be regularly distributed? Do the maples appear to be clustered or do they appear to be regularly distributed? Justify your answer. Simulation envelopes will help you give a better answer to this question.*

L transforms the K function by taking the square root of K. Under CSR, K is a parabola and by taking the square root of K (L) we can assess deviations from a straight line vs. a parabola when evaluating the CSR assumed hypothesis.

```r
par(mfrow=c(1,2))
plot(envelope(blackoak,fun="Lest", correction="iso", verbose = FALSE),
     main = "Blackoak")
plot(envelope(maple, fun="Lest", correction="iso", verbose = FALSE),
     main = "Maple")
```



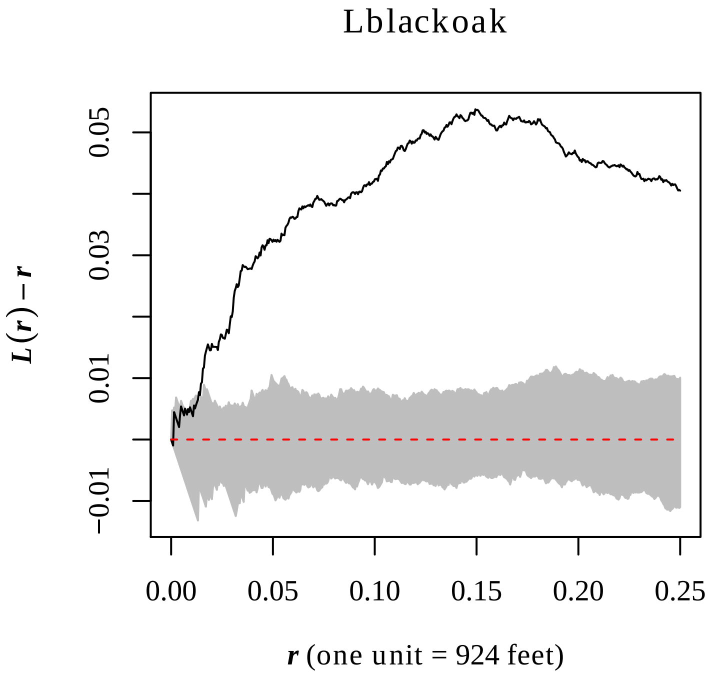
I used the isotrophic edge corrections. Both the blackoak and maple trees appear to be clustered as the observed number of additional events within almost all distances r is more than expected under CSR for both the blackoak and maple trees. Note that I plotted the simulation envelopes and the observed is outside of the simulation envelopes for most distances as well in both plots.

(c) *Compare the two L functions and discuss whether or not the 2 processes appear to be the same. You can use the results from (a) but you should also look at the difference more formally using the following also provided in an attached script file.*

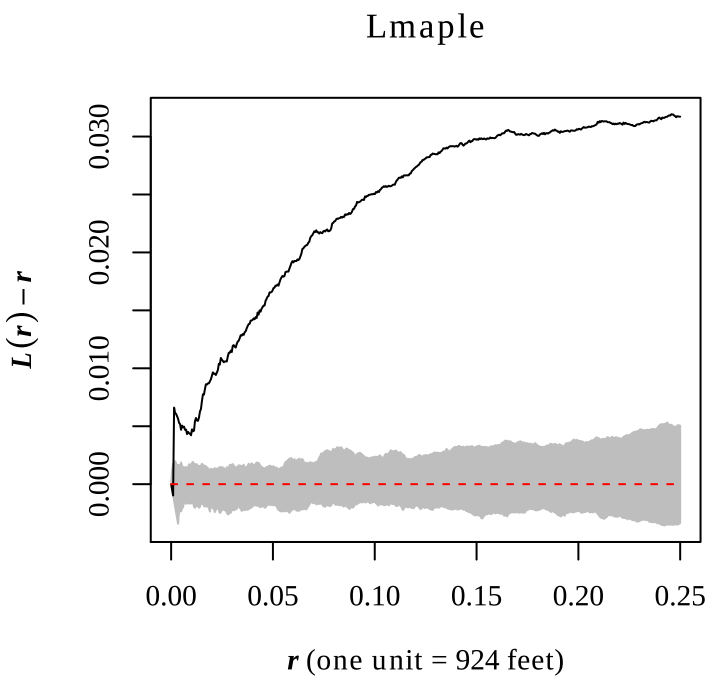
The two L functions do not appear to be the same as the D function is outside of the simulation envelopes for short distances.

The density plots shows the maple point process and the blackoack point process are near complements of each other, with different patterns and thus we might expect different L functions. Note the frequency of maples is much higher than that of the blackoak within the corresponding clusters.

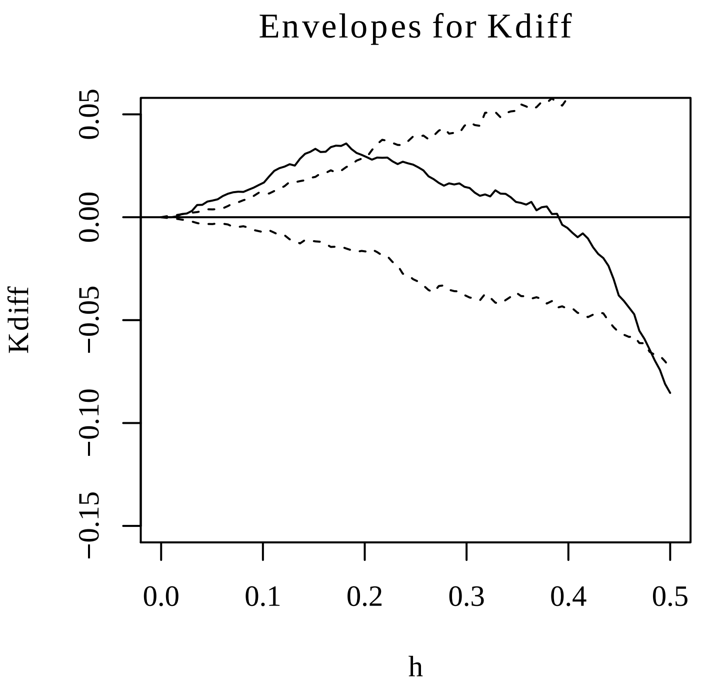
```
#r code embedded in midterm
#chose to output code from file
# specify radii
        h<-seq(0,.5,l=100)
        # get coordinates
        tree.poly<-list(x=c(blackoak$x,maple$x),y=c(blackoak$y,maple$y))
        # recompute the K functions
        kblackoak<-khat(as.points(blackoak),bboxx(bbox(as.points(tree.poly))),h)
        kmaple<-khat(as.points(maple),bboxx(bbox(as.points(tree.poly))),h)
        # get the differences
        k.diff<-kblackoak - kmaple
        # generate the envelope
        env<-Kenv.label(as.points(blackoak),as.points(maple),
        bboxx(bbox(as.points(tree.poly))),nsim=99,s=h, quiet = TRUE)
        # plot the results
        #plot(h,seq(-0.15,0.05,l=length(h)),type="n",ylab="Kdiff",
        #main="Envelopes for Kdiff")
        #lines(h,k.diff)
        #lines(h,env£low,lty=2)
        #lines(h,env£up,lty=2)
        #abline(h=0)

# r code sent as sepearate file
Lblackoak<-envelope(blackoak,fun=Lest,correction="iso")

Generating 99 simulations of CSR  ...
1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32
39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 6
77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98,  99.

Done.

plot(Lblackoak,.-r~r,legend=F)
```



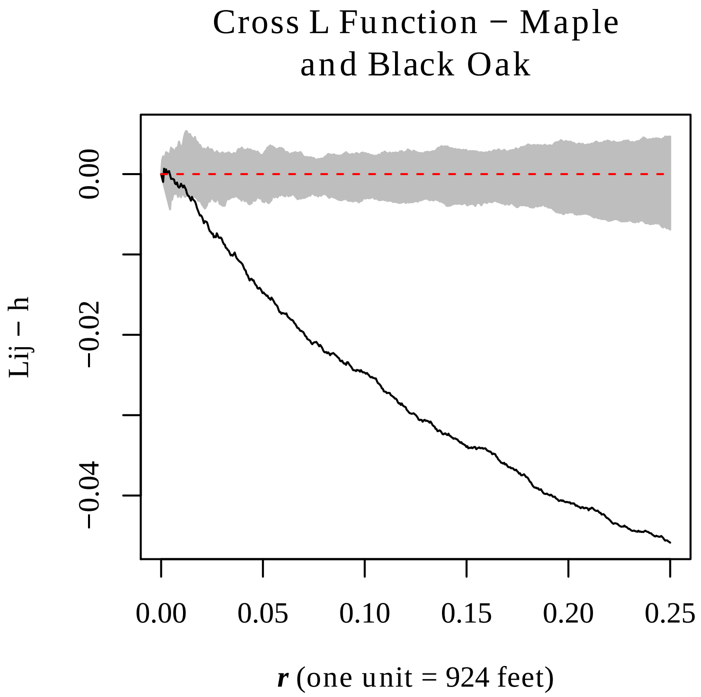
```
Lmaple<-envelope(maple,fun=Lest,correctin="iso")

Generating 99 simulations of CSR  ...
1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32
39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 6
77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98,  99.

Done.

plot(Lmaple,.-r~r,legend=F)
```



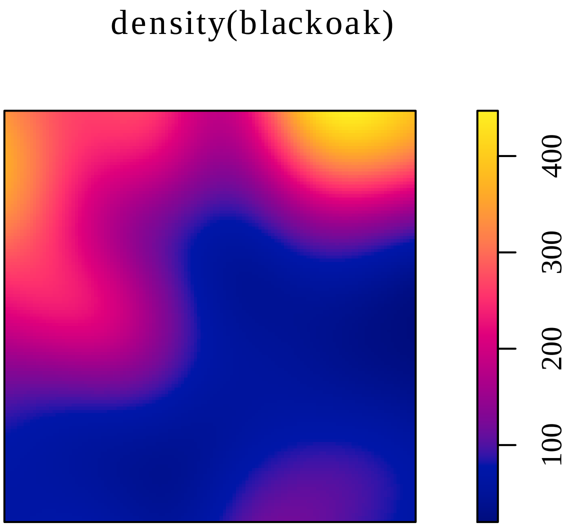
```
# specify radii
h<-seq(0,.5,l=100)
# get coordinates
tree.poly<-list(x=c(blackoak$x,maple$x),y=c(blackoak$y,maple$y))
# recompute the K functions
kblackoak<-khat(as.points(blackoak),bboxx(bbox(as.points(tree.poly))),h)
kmaple<-khat(as.points(maple),bboxx(bbox(as.points(tree.poly))),h)
# get the differences
k.diff<-kblackoak - kmaple
# generate the envelope
env<-(Kenv.label(as.points(blackoak),as.points(maple),
bboxx(bbox(as.points(tree.poly))),nsim=99,s=h, quiet = TRUE))
# plot the results
plot(h,seq(-0.15,0.05,l=length(h)),type="n",ylab="Kdiff",
main="Envelopes for Kdiff")
lines(h,k.diff)
lines(h,env$low,lty=2)
lines(h,env$up,lty=2)
abline(h=0)
```



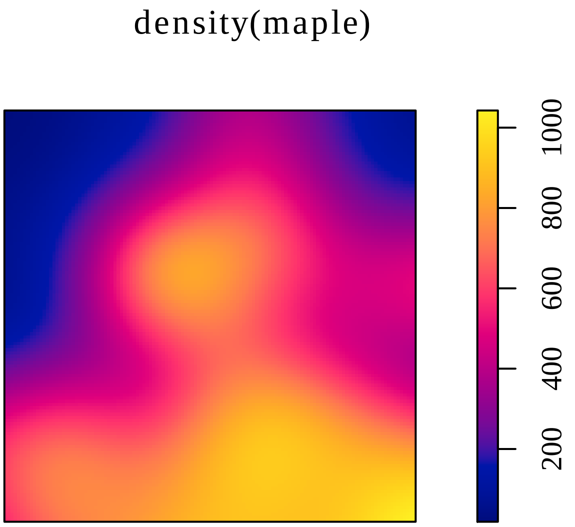
```
Kenv<-envelope(lansing,Kcross, i="maple",j="blackoak", verbose = FALSE)
plot(Kenv,sqrt(./pi)-r~r,ylab="Lij - h",main="Cross L Function - Maple
and Black Oak",legend=F)
```



```
plot(density(blackoak))
```

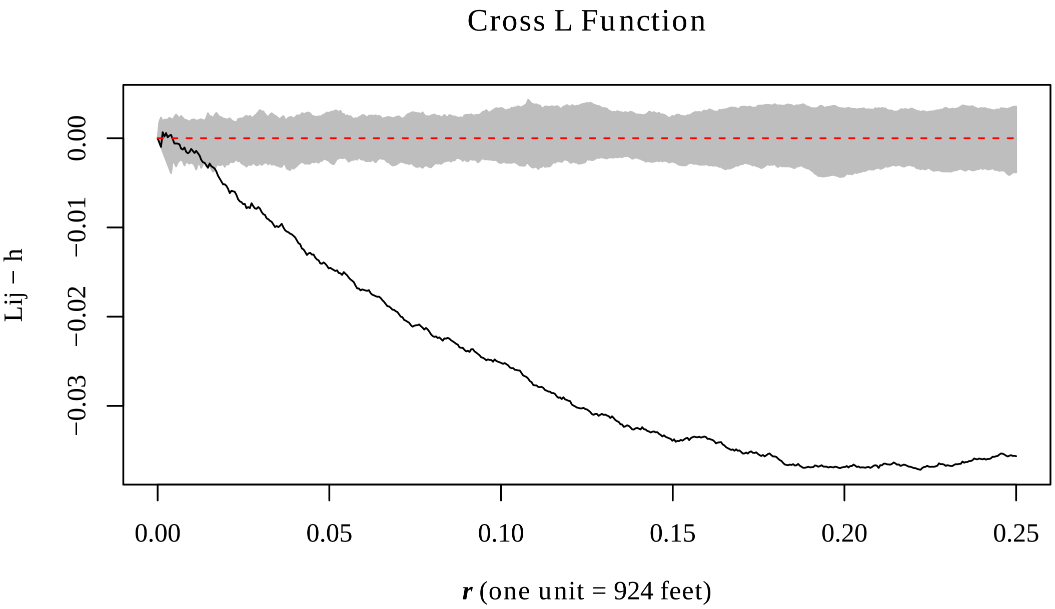


```
plot(density(maple))
```



(d) *Plot Lij h versus h for black oaks and maples.*

```
Kplot<-envelope(lansing,Kcross,i="blackoak",j="maple", verbose = FALSE)
plot(Kplot,sqrt(./pi)-r~r,ylab="Lij - h",main="Cross L Function",legend=F)
```



What type of relationship between the point patterns of the two species of trees is indicated by this plot? Justify your answer.

Because the observed Lij - h curve is below the theoretical (including below the corresponding simulation envelopes), the plot indicates the two species of trees are inhibiting the growth of each other, e.g., we do not expect to see many blackoaks near a cluster of maples.

(e) *Based on the above, comment on the null hypotheses of independence and random labeling.*

Independence H_o: The spatial locations and the binary marks are determed simultaneously and independently of one another, meaning as $K_{ij}(h)$ is defined in the notes at the top of page 38 can be simplified to $K(h) = \pi * h^2$. This implies the expected number of additional trees within distance h of one species of trees is the same regardless of which species is of interest and only depends on the distance considered, h.

The plot in (d) suggests evidence against the null hypothesis of independence and evidence for inhibition.

Mathematically this is violated because under the null hypothesis $E[K_{ij}(h)] =$ h.

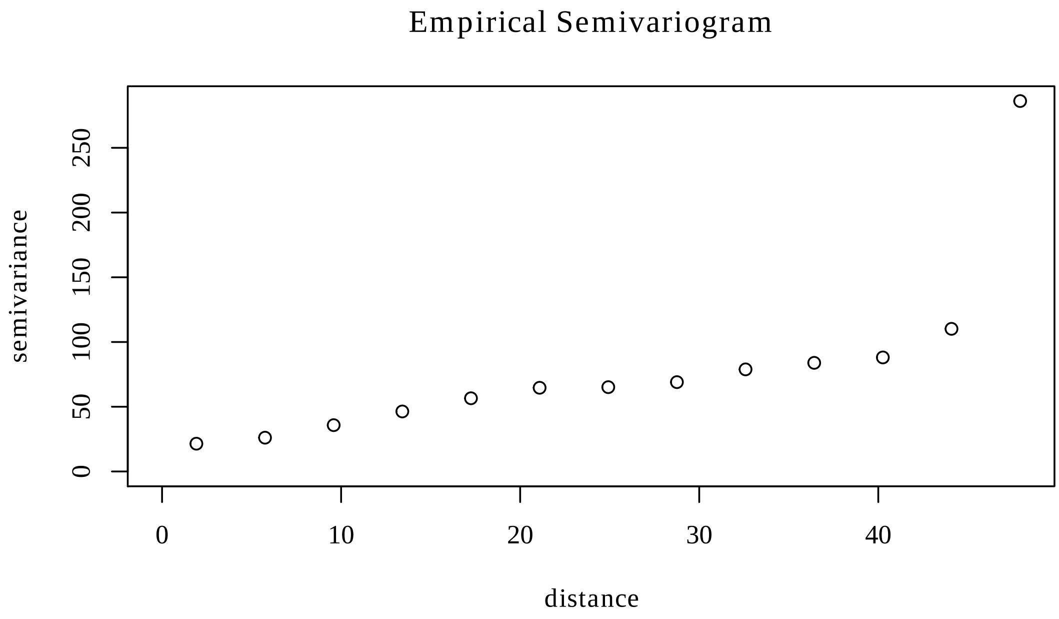
Random Labelling H_o: Locations arise from a univariate point process and the labels are determined by a process similar to random thinning, where the label is determined by a Bernoulli trial at each location. Where 1 indicates a blackoak and 2 indicates a maple, the null hypothesis corresponds to $K_{11} = K_{22} = K_{12} = K$.

The Kdiff plot in (c) suggests $K_{11} \neq K_{22}$, which should hold under random labelling. Therefore, random labelling of the process is also not suggested.

4. *You were sent the wheat data set on a previous homework assignment. You want to predict the value of Z (yield) at an arbitrary location. Assume a pure nugget effect model.*

(a) *What are the kriging weights and what is the predicted value?*

```
plot(wheat.variog <- variog(wheat.geodat),
     main = "Empirical Semivariogram")

variog: computing omnidirectional variogram
```



```
wheat.dist <- as.matrix(dist(wheat.geodat$coords))
d <- dim(wheat.dist)

# what is the nugget?
# using empirical right now
Sigma <- diag(var(wheat.geodat$data), nrow = d[1], ncol = d[1])

a <- c(rep(1,d[1]))

Sigma.star <- as.matrix(cbind((rbind(Sigma,a)),c(a,0)),
                        nrow = 225, ncol = 225)
str(Sigma.star)

 num [1:225, 1:225] 55.5 0 0 0 0 ...
 - attr(*, "dimnames")=List of 2
  ..$ : chr [1:225] "" "" "" "" ...
  ..$ : NULL
```

```
dim(Sigma.star)

[1] 225 225

newx <- runif(1,min(wheat.geodat$coords[,1]), max(wheat.geodat$coords[,1]))
newy <- runif(1,min(wheat.geodat$coords[,2]), max(wheat.geodat$coords[,2]))

sigma.vec <- as.matrix(dist(rbind(wheat.geodat$coords, c(newx,newy))))
sigma.vec <- sigma.vec[225,]

sigma.star <- c(rep(0,224),1)
lambda.star <- solve(Sigma.star) %*% sigma.star
lambda.star <- round(lambda.star,4)

p.ok <- sum(lambda.star[-225]*wheat.geodat$data)
# the predicted value is just the mean

# note in pure nugget cov.pars shouldn't matter
```

The predicted value is the mean, 25.731225, and all 224 of the weights are 0.0045.

(b) *What is the estimate of the sill?*

Note the sill is the nugget in a pure nugget model. The estimate of the sill is the estimate of the nugget, which was 55.5051753.

(c) *What is the kriging standard error (note that this is a* **prediction** *error)?*

```
c0 <- Sigma[1,1]

sigma2.ok <- c0 - sum(lambda.star*sigma.star)
# took sqrt of sigma2.ok for se
```

The kriging standard error is 7.4667915.

5. *Carbon-Nitrogen data example: We looked at estimating the semivariogram of the residuals from a simple linear regression model of total carbon on total nitrogen in class. We used gls to do this (as part of incorporating a spatial covariance structure into the regression) specifying an* **exponential** *covariance model and estimating the parameters using both* **maximum likelihood and REML**. *Let's check to see what the lik.fit function in the geoR package would return as parameter estimates (nugget, practical range, and partial sill) and see if the results are comparable. Some of the relevant R code is included in the attached script file. Compare the estimates on page 11 of the Spatial Regression notes and the estimates you get out of lik.fit. Use the same starting values.*

```
CN.dat <- read.table("CN.dat", header = TRUE)

pred.grid<-expand.grid(seq(-50,550,l=100),seq(-15,330,l=100))

TC.geodat<-as.geodata(CN.dat,coords.col=1:2,data.col=4)


# Problem 5

# get the CN data
names(CN.dat)<-c("x","y","tn","tc","cn")
```

```
#attach(CN.dat)
CN.lm<-lm(tc~tn, data = CN.dat)
resids<-residuals(CN.lm)
# convert to a geodata object
resids.dat<-cbind(CN.dat$x,CN.dat$y,resids)
resids.dat<-data.frame(resids.dat)
names(resids.dat)<-c("x","y","resids")
resids.geodat<-as.geodata(resids.dat,coords.col=1:2,data.col=3)


# now fit the models.

# gls initial params = c(range,nugget/sill)

# likfit initial params = c(sill-nugget, range/3)

# problem: NEED sill and nugget, not proportion
# look back at variog on p.9 to get corresponding
# starting values
# looks like sill is about 1.2
# this means nugget is about 1.2*0.4=0.48
# needed to do this to have similar starting values
# as those in the notes

range.init <- 15
nugget.init <- 0.48
sill.init <- 1.2

resids.ML <- likfit(resids.geodat, resids.geodat$coords,
                     resids.geodat$data, cov.model = "exponential",
                     ini.cov.pars = c(sill.init-nugget.init,range.init/3),
                    fix.nugget = FALSE, nugget = nugget.init,
                    lik.method = "ML", hessian = TRUE, messages = FALSE)

resids.REML <- likfit(resids.geodat, resids.geodat$coords,
                     resids.geodat$data, cov.model = "exponential",
                     ini.cov.pars = c(sill.init-nugget.init,range.init/3),
                    fix.nugget = FALSE, nugget = nugget.init,
                    lik.method = "REML", hessian = TRUE, messages = FALSE)

resids.ML$cov.pars

[1]  0.001065496 47.224029867

resids.ML$nugget

[1] 0.0005981685

resids.REML$cov.pars

[1]  0.001133522 56.242656929
```

REML

The table on p.11 of the notes using gls includes the proportion of the sill that is due to the nugget effect (estimated to be 0.356 with a confidence interval of 0.194 to 0.559). The gls function estimated the nugget effect to be 0.000621. Likfit estimated the nugget effect to be 6.2×10^{-4}, and so the results are similar.

Gls estimated the sill to be 0.00177 with a confidence interval ranging from 0.0013 to 0.00215. Likfit estimated the partial sill to be 0.00113 and thus the sill to be $0.00113 + 6.2 \times 10^{-4} = 0.00175$. Again, the likfit estimate falls within the confidence interval based on the gls estimate and the results between gls and likfit are comparable.

Gls estimated the practical range to be 175.089 with a confidence interval ranging from 58.49 to 524.131. Likfit estimated the practical range to be 3*56.24 = 168.73. We see again the results are comparable.

MAXIMUM LIKLIHOOD

Gls estimated the nugget to be $0.00167 * 0.363 = 6.06 \times 10^{-4}$ (proportion of the sill that is due to the nugget was estimated to be 0.363 with a confidence interval of 0.196 to 0.571). Likfit estimated the nugget to be 5.98×10^{-4}.

Gls estimated the sill to be 0.00167 with a confidence interval ranging from 0.00121 to 0.0029. Likfit estimated the partial sill to be 0.00107 and thus the sill to be $0.00107 + 6 \times 10^{-4} = 0.00166$.

Gls estimated the practical range to be 144.929 with a confidence interval ranging from 58.588 to 358.515. Likfit estimated the practical range to be 3*47.22 = 141.67.

We see again the results are comparable for all three parameters whether using likfit and ML or gls and ML.