Simulation is an important tool for research. Through simulation studies, we can evaluation a statistical method and enhance our understanding of statistical inferences. In this note, I will show you some basics of simulation study and assign to homework problems at the end.

1 Simulation of spatial data

Assume we simulate normal data. What we do is essentially to simulate from the multivariate normal distribution. The function `mvrnorm` in the `MASS` library can sample from a multivariate normal distribution. When the sample size is less than 6,000, you can performance such simulations in most of PCs.

```
cov.matern=function(x, nu = 2, alpha = 1, vars=1)
{
    if(nu == 0.5)
        return(vars*exp( - x * alpha))
    ismatrix <- is.matrix(x)
    if(ismatrix){nr=nrow(x); nl=ncol(x)}
    x <- c(alpha * x)
    output <- rep(1, length(x))
    n <- sum(x > 0)
    if(n > 0) {
        x1 <- x[x > 0]
        output[x > 0] <-
        (1/((2^(nu - 1)) * gamma(nu))) * (x1^nu) * besselK(x1, nu)
    }
    if(ismatrix){
        output <- matrix(output, nr, nl)
        }
    vars*output
}
locs=cbind(rep(0:20, 21)/20, rep(0:20, each=21)/20)
V=cov.matern(as.matrix(dist(locs)),nu=1/2, alpha=7*sqrt(3))
set.seed(20)
z=mvrnorm(mu=rep(0, 21^2), Sigma=V)
z=matrix(z, ncol=21)
persp(x=(0:20)/21, y=(0:20)/21, z,  theta=45, phi=35, r=5, expand=0.6, axes=T,
ticktype="detailed", xlab="x", ylab="y", zlab="z")
filled.contour(x=0:20, y=0:20, z, color.palette=gray.colors)
```

```
V=cov.matern(as.matrix(dist(locs)),nu=2, alpha=7*sqrt(3))
set.seed(20)
z=mvrnorm(mu=rep(0, 21^2), Sigma=V)
z=matrix(z, ncol=21)
```

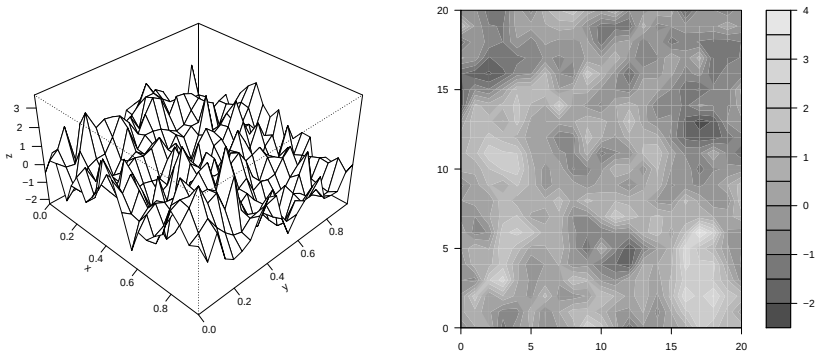


Figure 1: Simulated surface with Matern $\nu = 1/2$

```
persp(x=(0:20)/21, y=(0:20)/21, z,  theta=45, phi=35, r=5, expand=0.6, axes=T,
ticktype="detailed", xlab="x", ylab="y", zlab="z")
filled.contour(x=0:20, y=0:20, z, color.palette=gray.colors)
```

```
V=cov.matern(as.matrix(dist(locs)),nu=10, alpha=7*sqrt(3))
set.seed(20)
z=mvrnorm(mu=rep(0, 21^2), Sigma=V)
z=matrix(z, ncol=21)
persp(x=(0:20)/21, y=(0:20)/21, z,  theta=45, phi=35, r=5, expand=0.6, axes=T,
ticktype="detailed", xlab="x", ylab="y", zlab="z")
filled.contour(x=0:20, y=0:20, z, color.palette=gray.colors)
```

2 Repeated Simulation and Estimation

Often we need to repeatedly simulate independent copies of random variables. In this situation, we need to store the results in each simulation.

Example 1. We know from the Central Limit Theorem (CLT) that the sample mean $\bar{X}_n = (1/n)\sum_{i=1}^{n} X_i$ is appropriately normal when n is large. However, the CLT does not tell us how large n should be in order for $\bar{X}_n$ to be close to normal. We will use simulation to find out the sampling distribution of $\bar{X}_n$ when $X_1, \ldots, X_n$ are i.i.d. U[0,1].

```
n=20 #n is the sample size
N=200 # N is the simulation size
result=mat.or.vec(N, 1)
```

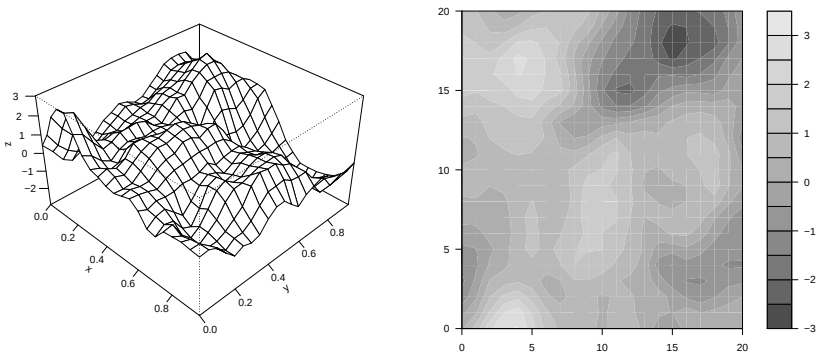


Figure 2: Simulated surface with Matern $\nu = 2$

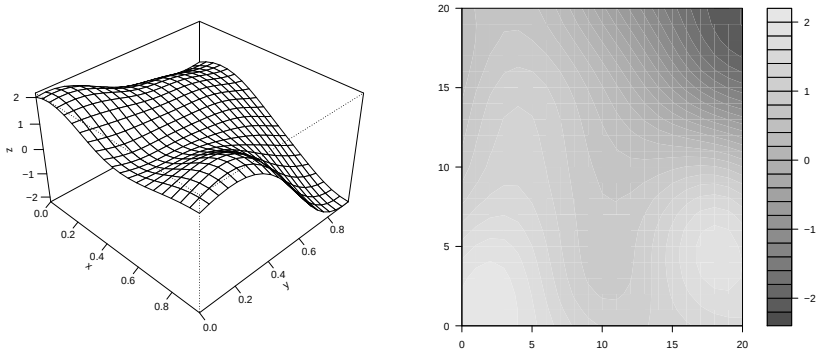
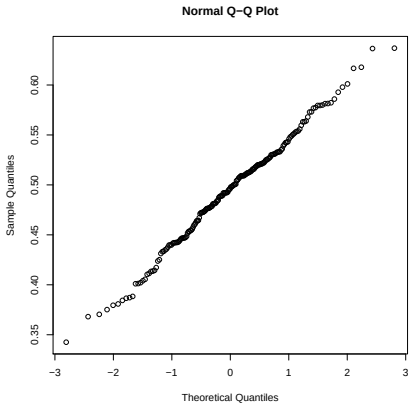


Figure 3: Simulated surface with Matern $\nu = 10$

```
i=0
repeat{
i=i+1
if(i>N) break
result[i]=mean(runif(n))
}
qqnorm(result)
```



Example 2. We now use simulation to find out the sampling distribution of $\bar{X}_n$ when $X_1, \ldots, X_n$ are i.i.d. U[0,1].

```
n=20 #n is the sample size
N=200 # N is the simulation size
result=mat.or.vec(N, 1)
i=0
repeat{
i=i+1
if(i>N) break
result[i]=mean(rexp(n, 1))
}
qqnorm(result)
```

We see that $n = 20$ is not sufficiently large. We now incease it to 40

```
n=40 #n is the sample size
N=200 # N is the simulation size
result=mat.or.vec(N, 1)
i=0
repeat{
i=i+1
if(i>N) break
result[i]=mean(rexp(n,1))
```

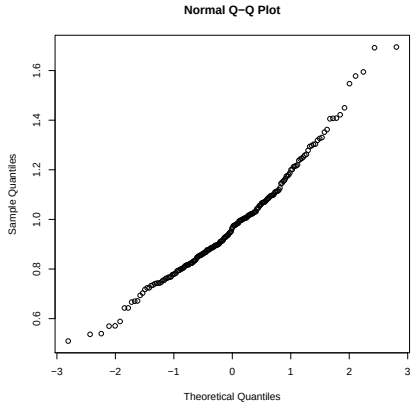


Figure 4: QQ plot of sample means given i.i.d sample from Exp(1) distribution with sample size $n = 20$.

```
}
qqnorm(result)
```

3 Project

This project is due March 27.

1. You will run some simulations to replicate the results in Zhang (2004) by following the following steps.

   (a) Define the 221 sampling locations where your simulated data are observed: $(i/10, j/10), i, j = 0, 1, \cdots, 10$ and $\{(0.05 + 0.1i, 0.05 + 0.1j), i, j = 0, \cdots, 9\}$. You should put these 221 locations into a 221×2 matrix and name it, say, *locs*.

   (b) Use the exponential covariogram $K_0(x) = \sigma_0^2 \exp(-x/\theta_0)$, $x \geq 0$, where $\sigma_0^2 = 1$ and $\theta_0 = 0.2$ to calculate the covariance matrix V of the 221 random variables at the sampling locations.

   (c) Use the function *mvrnorm* in the R library *MASS* to generate 221 normal random variables at the 221 locations corresponding to the covariance matrix V (assume the mean is known to be 0).

   (d) Use the simulated data and the covariogram K_0 to calculate the simple kriging prediction at 31 locations $(0.387, 0.1 + 0.01n)$, $n = 0, \cdots, 30$. Record the predicted values and the prediction variances.

   (e) Now repeat (2d) twice by using the covariograms $K_1(x) = 2\exp(-x/0.4)$ and $K_2 = 1.8\exp(-x/0.4)$. Record the predicted values and the prediction variances for each covariogram.

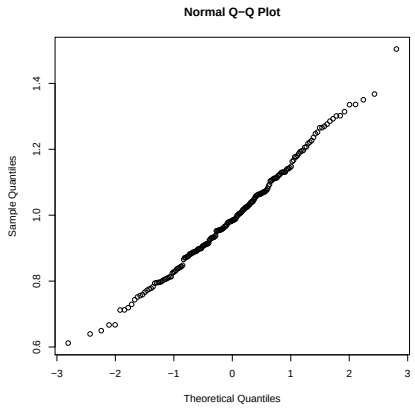


Figure 5: QQ plot of sample means given i.i.d sample from Exp(1) distribution with sample size $n = 40$.

(f) Plot the 3 sets of predicted value against the prediction locations in one plot.

(g) Plot the 3 sets of prediction variance against the prediction locations in one plot.

(h) What can you conclude from the plots? Can you give any theoretical justification?

2. In this simulation study, you will investigate the effects of tapering on the prediction.

(a) Define the 221 sampling locations where your simulated data are observed: $(i/10, j/10), i, j = 0, 1, \cdots, 10$ and $\{(0.05 + 0.1i, 0.05 + 0.1j), i, j = 0, \cdots, 9\}$. You should put these 221 locations into a 221×2 matrix and name it, say, *locs*.

(b) Use the exponential covariogram $K_0(x) = \sigma_0^2 \exp(-x/\theta_0),\ x \geq 0$, where $\sigma_0^2 = 1$ and $\theta_0 = 0.2$ to calculate the covariance matrix V of the 221 random variables at the sampling locations.

(c) Use the function *mvrnorm* in the R library *MASS* to generate 221 normal random variables at the 221 locations corresponding to the covariance matrix V (assume the mean is known to be 0).

(d) Use the simulated data and the covariogram K_0 to calculate the simple kriging prediction at 31 locations $(0.387, 0.1 + 0.01n)$, $n = 0, \cdots, 30$. Record the predicted values and the prediction variances.

(e) Now repeat (2d) twice by using the covariograms

$$K_1(h) = 2\exp(-h/0.4)(1-h/0.2)_+^6(1+6(h/0.2)+(35/3)(h/0.2)^2), h \geq 0$$

and

$$K_2(h) = 2\exp(-h/0.4)(1-h/0.3)_+^6(1+6(h/0.3)+(35/3)(h/0.3)^2), h \geq 0.$$

where a_+ denotes $\max(0, a)$ for any number a. Record the predicted values and the prediction variances for each covariogram.

(f) Plot the 3 sets of predicted value against the prediction locations in one plot.

(g) Plot the 3 sets of prediction variance against the prediction locations in one plot.

(h) What can you conclude from the plots?

References

Zhang, H. (2004). Inconsistent estimation and asymptotically equivalent interpolations in model-based geostatistics. *Journal of the American Statistical Association*, **99**, 250–261.