Real Data Analysis Final

Li Chen

5/30/2022

Contents


```
load("gene_expr_data.RData")
source("../mbtlp.R",chdir = TRUE)
```

```
## Loading required package: lattice
```

```
## Loading required package: robustbase
```

```
source("../intdagdiscovery.R",chdir = TRUE)
source("../intdagcoef.r",chdir = TRUE)
source("../intdaginfer.r",chdir = TRUE)
```

```r
is.acyclic <- function(U){
  flag <- 1
  while (sum(U)>0){
    if (min(colSums(U))>0){
      flag <- 0
      break
    }
    idx <- which(colSums(U)!=0)
    U <- U[idx,idx,drop=FALSE]
  }
  return(flag)
}
```

Preprocessing

```r
p <- 146
q <- 2*p
a <- rep(NA,p)

snp.AD <- rep(NA,p)
snp.CN <- rep(NA,p)
for(j in 1:p) {
  m1 <- lm(Y.CN[,j] ~ X.CN[,2*j-1])
  m2 <- lm(Y.CN[,j] ~ X.CN[,2*j])
  str1.CN <- summary(m1)$coefficients[,4]
  str2.CN <- summary(m2)$coefficients[,4]
  str.CN <- min(str1.CN[-1],str2.CN[-1])
  snp.CN[j] <- str.CN

  m1 <- lm(Y.AD[,j] ~ X.AD[,2*j-1])
  m2 <- lm(Y.AD[,j] ~ X.AD[,2*j])
  str1.AD <- summary(m1)$coefficients[,4]
  str2.AD <- summary(m2)$coefficients[,4]
  str.AD <- min(str1.AD[-1],str2.AD[-1])
  snp.AD[j] <- str.AD

  a[j] <- min(str.AD, str.CN)
}


gene <- which(a < 1e-14) # change the threshold for significant SNPs

snp <- rep(0, length(gene))
for (k in 1:length(gene)) {
  snp[2*k-1] <- 2*gene[k]-1
  snp[2*k] <- 2*gene[k]
}

gene_name <- names(Y.AD)[gene]
gene_name
```

```
##  [1] "APBB1"    "APOE"     "APP"      "ATP5F1"   "CAPN1"    "CAPN2"
##  [7] "CASP3"    "CASP8"    "CDK5R1"   "COX7A2L"  "FADD"     "GAPDH"
```

```
## [13] "GSK3B"    "ITPR2"    "LPL"      "LRP1"     "NDUFA2"   "NDUFV3"
## [19] "PSEN1"    "SDHC"     "TNFRSF1A"
```

```
snp_name <- colnames(X.AD)[snp]
snp_name
```

```
##  [1] "rs10769692"  "rs2075583"   "rs78986976"  "rs11667253"  "rs4817078"
##  [6] "rs114233663" "rs12752970"  "rs1264898"   "rs12422027"  "rs1195968"
## [11] "rs59990581"  "rs41267355"  "rs72689214"  "rs4862384"   "rs1035142"
## [16] "rs700636"    "rs1018866"   "rs3814984"   "rs1981664"   "rs12712839"
## [21] "rs1317742"   "rs1131715"   "rs1803621"   "rs3741918"   "rs1488763"
## [26] "rs62266319"  "rs75404742"  "rs4964018"   "rs78299715"  "rs80073370"
## [31] "rs7975818"   "rs7489208"   "rs2563293"   "rs1962649"   "rs4148974"
## [36] "rs2839603"   "rs214267"    "rs177394"    "rs16832846"  "rs56871324"
## [41] "rs4149576"   "rs2302350"
```

AD part

```
Y1 <- Y.AD[,gene]
X1 <- X.AD[,snp]
```

```
Y1 <- t(t(Y1)-colMeans(Y1))
X1 <- t(t(X1)-colMeans(X1))
```

Estimation of V matrix

```
set.seed(0)
tau.list <- c(0.01,0.02,0.03)
gamma.list <- seq(0.00001,0.001,0.00001)
n.fold <- 5
result1.1 <- cv.intdag.pmle.diff.aic(X1,Y1,tau.list,gamma.list,n.fold)
```

Structure Recovery

```
V <- result1.1$V
result1.2 <- topological_order(V)
```

Coefficient estimation

```
Pi1 <- result1.2$an_mat
Phi1 <- result1.2$in_mat
Piv1 <- result1.2$iv_mat
```

```
set.seed(0)
n.fold <- 5
tau.list <- c(0.01,0.02,0.03)
gamma.list <- seq(0.1,3.5,0.1)
result1.3 <- cv.intdag.coe(X1,Y1,Pi1,Phi1,Piv1,tau.list,gamma.list,n.fold)
```

mbtlp for precision matrix estimation

```
set.seed(0)
Z1 <- Y1 - Y1%*%result1.3$U - X1%*%result1.3$W
```

```
tau.list <- c(0.01,0.02,0.03)
gamma.list <- seq(0,0.0001,0.000001)
n.fold <- 5
result1.5 <- cv.MB_Union(Z1,tau.list,gamma.list,n.fold)
```

```
S <- result1.5$S
Sigma <- result1.3$Sigma
max.it <- 10000
tol <- 1e-7
wi1 <- precision_refit(Sigma,S,max.it,tol)
```

APP -> APOE

```
idx1 <- which(colnames(Y1) == 'APP')
idx2 <- which(colnames(Y1) == 'APOE')
```

```
U.test <- matrix(0,nrow(Pi1),ncol(Pi1))
U.test[idx1,idx2] <- 1
U1 <- 1*((Pi1+U.test)>0)
is.acyclic(U1)==0
```

```
## [1] FALSE
```

```
U0 <- U1-U.test
```

```
stat1.1 <- intdag.2lr(X1,Y1,U0,Phi1,U1,Phi1,wi1)$statistic
stat1.1
```

```
## [1] 17.57762
```

LRP1 -> CASP3

```
idx1 <- which(colnames(Y1) == 'LRP1')
idx2 <- which(colnames(Y1) == 'CASP3')
```

```
U.test <- matrix(0,nrow(Pi1),ncol(Pi1))
U.test[idx1,idx2] <- 1
U1 <- 1*((Pi1+U.test)>0)
is.acyclic(U1)==0
```

```
## [1] FALSE
```

```
U0 <- U1-U.test
```

```
stat1.2 <- intdag.2lr(X1,Y1,U0,Phi1,U1,Phi1,wi1)$statistic
stat1.2
```

```
## [1] 50.36417
```

APP -> APBB1

```
idx1 <- which(colnames(Y1) == 'APP')
idx2 <- which(colnames(Y1) == 'APBB1')
```

```
U.test <- matrix(0,nrow(Pi1),ncol(Pi1))
U.test[idx1,idx2] <- 1
```

```
U1 <- 1*((Pi1+U.test)>0)
is.acyclic(U1)==0
```

```
## [1] FALSE
```

```
U0 <- U1-U.test
```

```
stat1.3 <- intdag.2lr(X1,Y1,U0,Phi1,U1,Phi1,wi1)$statistic
stat1.3
```

```
## [1] 52.73806
```

CAPN1 -> CDK5R1

```
idx1 <- which(colnames(Y1) == 'CAPN1')
idx2 <- which(colnames(Y1) == 'CDK5R1')
```

```
U.test <- matrix(0,nrow(Pi1),ncol(Pi1))
U.test[idx1,idx2] <- 1
U1 <- 1*((Pi1+U.test)>0)
is.acyclic(U1)==0
```

```
## [1] FALSE
```

```
U0 <- U1-U.test
```

```
stat1.4 <- intdag.2lr(X1,Y1,U0,Phi1,U1,Phi1,wi1)$statistic
stat1.4
```

```
## [1] 96.27275
```

LRP1 -> GSK3B

```
idx1 <- which(colnames(Y1) == 'LRP1')
idx2 <- which(colnames(Y1) == 'GSK3B')
```

```
U.test <- matrix(0,nrow(Pi1),ncol(Pi1))
U.test[idx1,idx2] <- 1
U1 <- 1*((Pi1+U.test)>0)
is.acyclic(U1)==0
```

```
## [1] FALSE
```

```
U0 <- U1-U.test
```

```
stat1.5 <- intdag.2lr(X1,Y1,U0,Phi1,U1,Phi1,wi1)$statistic
stat1.5
```

```
## [1] 0.04845182
```

CAPN1 -> CASP3

```
idx1 <- which(colnames(Y1) == 'CAPN1')
idx2 <- which(colnames(Y1) == 'CASP3')
```

```
U.test <- matrix(0,nrow(Pi1),ncol(Pi1))
U.test[idx1,idx2] <- 1
U1 <- 1*((Pi1+U.test)>0)
is.acyclic(U1)==0
```

```
## [1] FALSE
```

```
U0 <- U1-U.test
```

```
stat1.6 <- intdag.2lr(X1,Y1,U0,Phi1,U1,Phi1,wi1)$statistic
stat1.6
```

```
## [1] 0.3303143
```

ATP5F1 -> CASP3

```
idx1 <- which(colnames(Y1) == 'ATP5F1')
idx2 <- which(colnames(Y1) == 'CASP3')
```

```
U.test <- matrix(0,nrow(Pi1),ncol(Pi1))
U.test[idx1,idx2] <- 1
U1 <- 1*((Pi1+U.test)>0)
is.acyclic(U1)==0
```

```
## [1] FALSE
```

```
U0 <- U1-U.test
```

```
stat1.7 <- intdag.2lr(X1,Y1,U0,Phi1,U1,Phi1,wi1)$statistic
stat1.7
```

```
## [1] 0.007244571
```

CN Part

```
Y2 <- Y.CN[,gene]
X2 <- X.CN[,snp]
```

```
Y2 <- t(t(Y2)-colMeans(Y2))
X2 <- t(t(X2)-colMeans(X2))
```

Estimation of V matrix

```
set.seed(0)
tau.list <- c(0.01,0.02,0.03)
gamma.list <- seq(0.00001,0.001,0.00001)
n.fold <- 5
result2.1 <- cv.intdag.pmle.diff.aic(X2,Y2,tau.list,gamma.list,n.fold)
```

Structure Recovery

```
V <- result2.1$V
result2.2 <- topological_order(V)
```

Coefficient estimation

```
Pi2 <- result2.2$an_mat
Phi2 <- result2.2$in_mat
Piv2 <- result2.2$iv_mat
```

```
set.seed(0)
n.fold <- 5
tau.list <- c(0.01,0.02,0.03)
gamma.list <- seq(0.1,3.5,0.1)
result2.3 <- cv.intdag.coe(X2,Y2,Pi2,Phi2,Piv2,tau.list,gamma.list,n.fold)
```

mbtlp for precision matrix estimation

```
set.seed(0)
Z2 <- Y2 - Y2%*%result2.3$U - X2%*%result2.3$W
tau.list <- c(0.01,0.02,0.03)
gamma.list <- seq(0,0.0001,0.000001)
n.fold <- 5
result2.5 <- cv.MB_Union(Z2,tau.list,gamma.list,n.fold)
```

```
S <- result2.5$S
Sigma <- result2.3$Sigma
max.it <- 10000
tol <- 1e-7
wi2 <- precision_refit(Sigma,S,max.it,tol)
```

APP -> APOE

```
idx1 <- which(colnames(Y2) == 'APP')
idx2 <- which(colnames(Y2) == 'APOE')
```

```
U.test <- matrix(0,nrow(Pi2),ncol(Pi2))
U.test[idx1,idx2] <- 1
U1 <- 1*((Pi2+U.test)>0)
is.acyclic(U1)==0
```

```
## [1] FALSE
```

```
U0 <- U1-U.test
```

```
stat2.1 <- intdag.2lr(X2,Y2,U0,Phi2,U1,Phi2,wi2)$statistic
stat2.1
```

```
## [1] 0.01013474
```

LRP1 -> CASP3

```
idx1 <- which(colnames(Y2) == 'LRP1')
idx2 <- which(colnames(Y2) == 'CASP3')
```

```
U.test <- matrix(0,nrow(Pi2),ncol(Pi2))
U.test[idx1,idx2] <- 1
U1 <- 1*((Pi2+U.test)>0)
is.acyclic(U1)==0
```

```
## [1] FALSE
```

```
U0 <- U1-U.test
```

```
stat2.2 <- intdag.2lr(X2,Y2,U0,Phi2,U1,Phi2,wi2)$statistic
stat2.2
```

```
## [1] 3.475277
```

APP -> APBB1

```
idx1 <- which(colnames(Y2) == 'APP')
idx2 <- which(colnames(Y2) == 'APBB1')
```

```
U.test <- matrix(0,nrow(Pi2),ncol(Pi2))
U.test[idx1,idx2] <- 1
U1 <- 1*((Pi2+U.test)>0)
is.acyclic(U1)==0
```

```
## [1] FALSE
```

```
U0 <- U1-U.test
```

```
stat2.3 <- intdag.2lr(X2,Y2,U0,Phi2,U1,Phi2,wi2)$statistic
stat2.3
```

```
## [1] 9.98753
```

CAPN1 -> CDK5R1

```
idx1 <- which(colnames(Y2) == 'CAPN1')
idx2 <- which(colnames(Y2) == 'CDK5R1')
```

```
U.test <- matrix(0,nrow(Pi2),ncol(Pi2))
U.test[idx1,idx2] <- 1
U1 <- 1*((Pi2+U.test)>0)
is.acyclic(U1)==0
```

```
## [1] FALSE
```

```
U0 <- U1-U.test
```

```
stat2.4 <- intdag.2lr(X2,Y2,U0,Phi2,U1,Phi2,wi2)$statistic
stat2.4
```

```
## [1] 138.8592
```

LRP1 -> GSK3B

```
idx1 <- which(colnames(Y2) == 'LRP1')
idx2 <- which(colnames(Y2) == 'GSK3B')
```

```
U.test <- matrix(0,nrow(Pi2),ncol(Pi2))
U.test[idx1,idx2] <- 1
U1 <- 1*((Pi2+U.test)>0)
is.acyclic(U1)==0
```

```
## [1] FALSE
```

```
U0 <- U1-U.test
```

```
stat2.5 <- intdag.2lr(X2,Y2,U0,Phi2,U1,Phi2,wi2)$statistic
stat2.5
```

```
## [1] 4.327487
```

CAPN1 -> CASP3

```
idx1 <- which(colnames(Y2) == 'CAPN1')
idx2 <- which(colnames(Y2) == 'CASP3')
```

```
U.test <- matrix(0,nrow(Pi2),ncol(Pi2))
U.test[idx1,idx2] <- 1
U1 <- 1*((Pi2+U.test)>0)
is.acyclic(U1)==0
```

```
## [1] FALSE
```

```
U0 <- U1-U.test
```

```
stat2.6 <- intdag.2lr(X2,Y2,U0,Phi2,U1,Phi2,wi2)$statistic
stat2.6
```

```
## [1] 8.499405
```

ATP5F1 -> CASP3

```
idx1 <- which(colnames(Y2) == 'ATP5F1')
idx2 <- which(colnames(Y2) == 'CASP3')
```

```
U.test <- matrix(0,nrow(Pi2),ncol(Pi2))
U.test[idx1,idx2] <- 1
U1 <- 1*((Pi2+U.test)>0)
is.acyclic(U1)==0
```

```
## [1] FALSE
```

```
U0 <- U1-U.test
```

```
stat2.7 <- intdag.2lr(X2,Y2,U0,Phi2,U1,Phi2,wi2)$statistic
stat2.7
```

```
## [1] 61.73215
```

Summary of Results

```
stat1 <- c(stat1.1,stat1.2,stat1.3,stat1.4,stat1.5,stat1.6,stat1.7)
stat2 <- c(stat2.1,stat2.2,stat2.3,stat2.4,stat2.5,stat2.6,stat2.7)
stat.mat <- cbind(stat1,stat2)
colnames(stat.mat) <- c("AD","CN")
rownames(stat.mat) <- c("APP -> APOE","LRP1 -> CASP3","APP -> APBB1","CAPN1 -> CDK5R1","LRP1 -> GSK3B","
```

```
knitr::kable(stat.mat)
```

	AD	CN
APP -> APOE	17.5776202	0.0101347
LRP1 -> CASP3	50.3641652	3.4752767
APP -> APBB1	52.7380587	9.9875299
CAPN1 -> CDK5R1	96.2727546	138.8592227
LRP1 -> GSK3B	0.0484518	4.3274874
CAPN1 -> CASP3	0.3303143	8.4994047
ATP5F1 -> CASP3	0.0072446	61.7321453

Bonferroni-Holm Correction for linkage-test

```
stat.mat.correct <- stat.mat
ps  <- as.vector(stat.mat.correct)
ps <- unlist(lapply(ps,function(o) {return(1-pchisq(o,df=1))}))
p.correct <- p.adjust(ps,"holm")
p.mat.correct <- matrix(p.correct,ncol=2,byrow = FALSE)
colnames(p.mat.correct) <- colnames(stat.mat.correct)
rownames(p.mat.correct) <- rownames(stat.mat.correct)
```

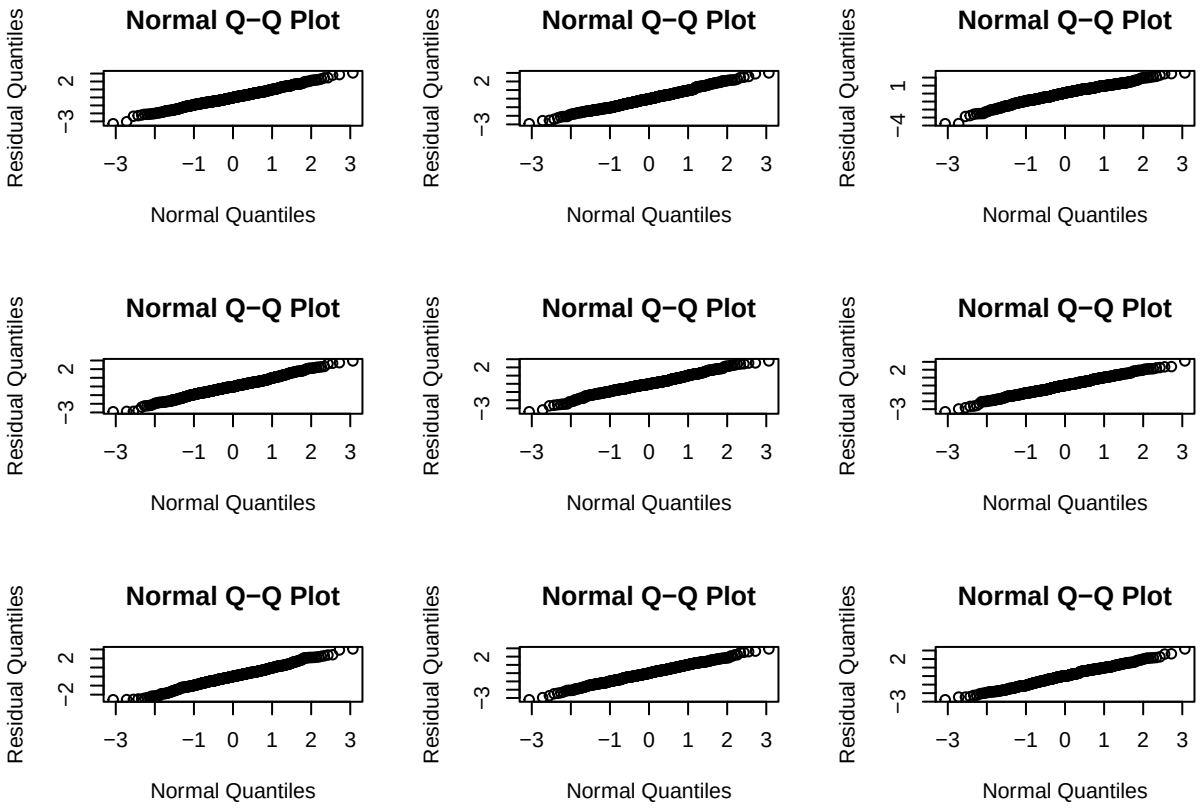
```
knitr::kable(p.mat.correct)
```

	AD	CN
APP -> APOE	0.0002482	1.0000000
LRP1 -> CASP3	0.0000000	0.3114615
APP -> APBB1	0.0000000	0.0126083
CAPN1 -> CDK5R1	0.0000000	0.0000000
LRP1 -> GSK3B	1.0000000	0.2250094
CAPN1 -> CASP3	1.0000000	0.0248684
ATP5F1 -> CASP3	1.0000000	0.0000000

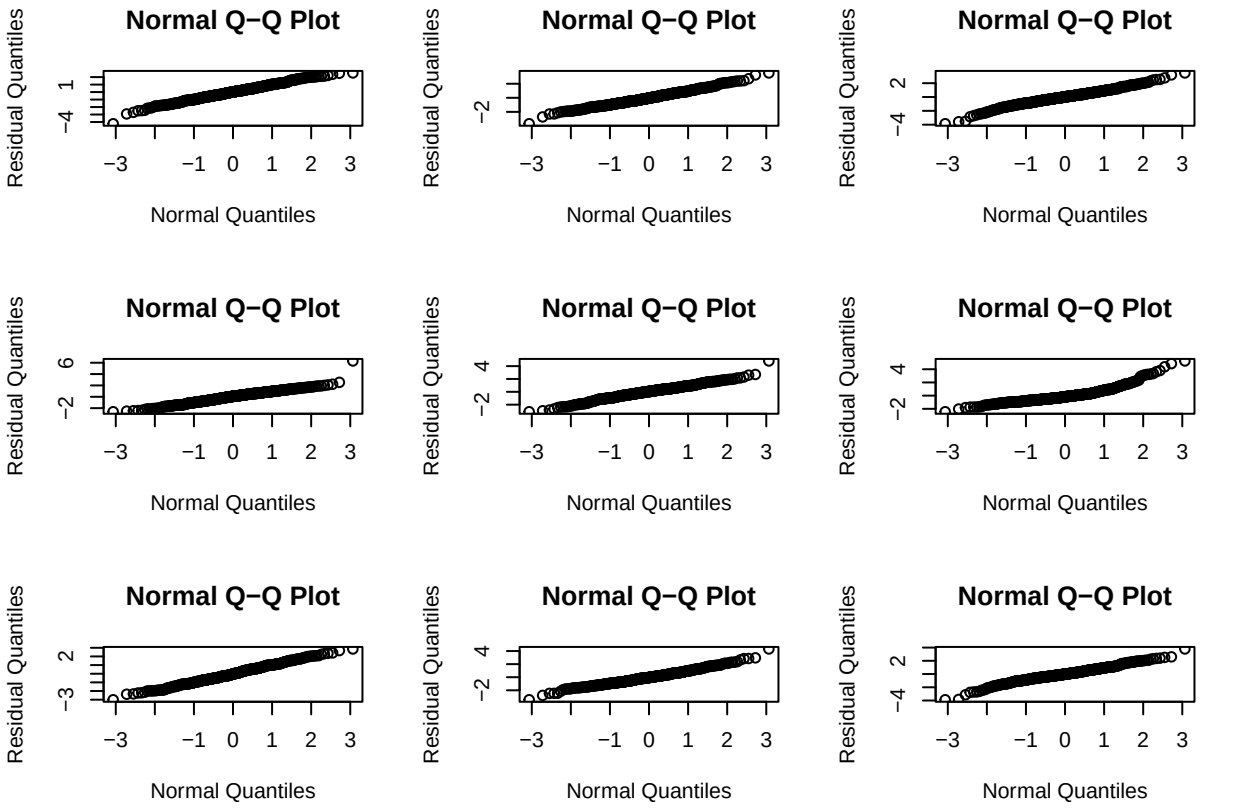
Normality Check

AD

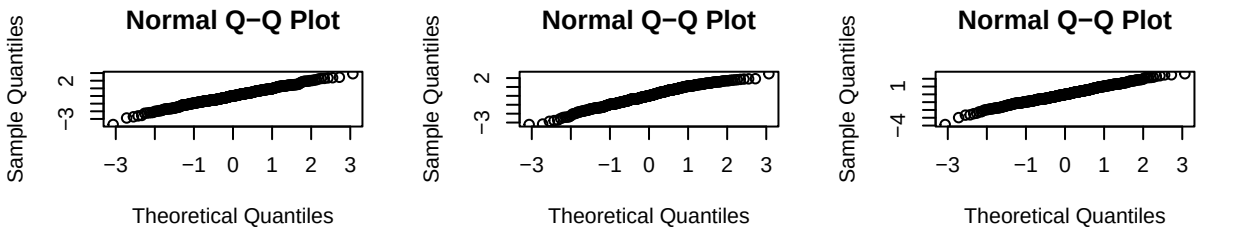
```
rmat1 <- apply(Z1,2,function(o){return(o/sd(o))})
par(mfrow=c(3,3))
for (i in 1:9){
  qqnorm(rmat1[,i],xlab = "Normal Quantiles", ylab = "Residual Quantiles")
}
```




```
par(mfrow=c(3,3))
for (i in 10:18){
  qqnorm(rmat1[,i],xlab = "Normal Quantiles", ylab = "Residual Quantiles")
}
```

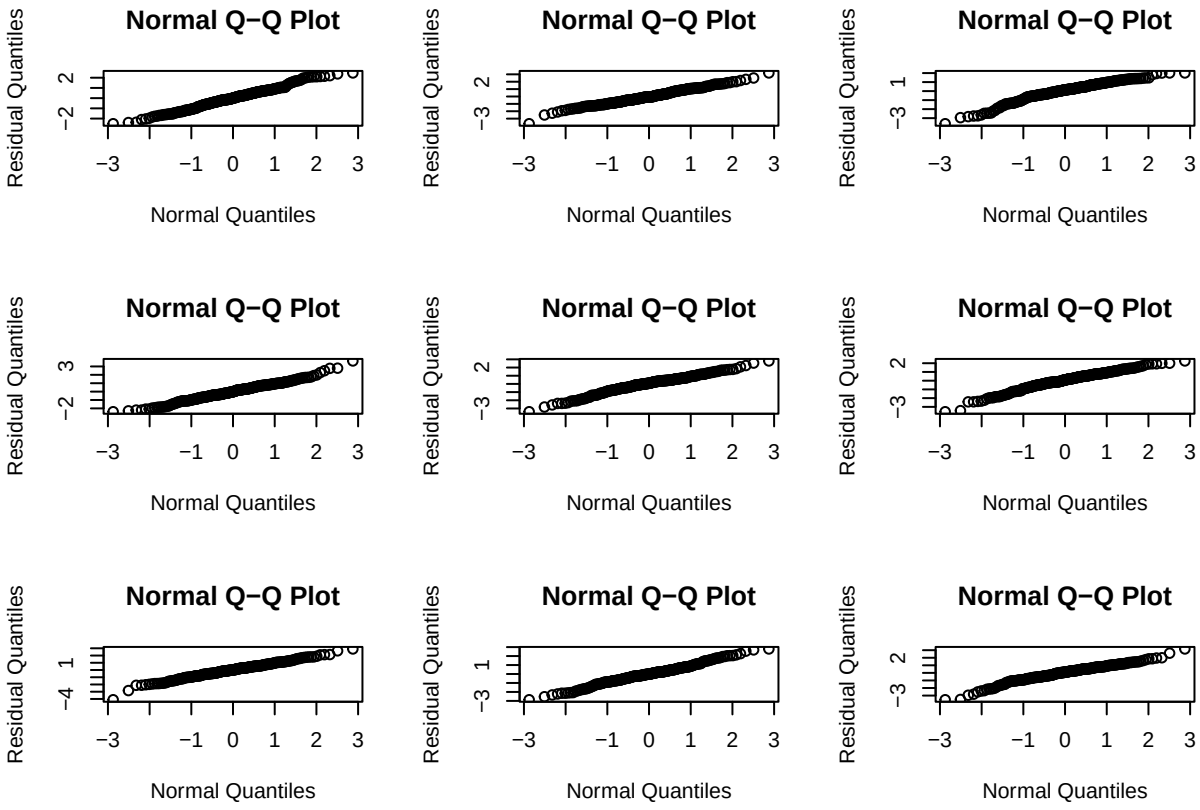



```
par(mfrow=c(3,3))
for (i in 19:21){
  qqnorm(rmat1[,i])
}
```

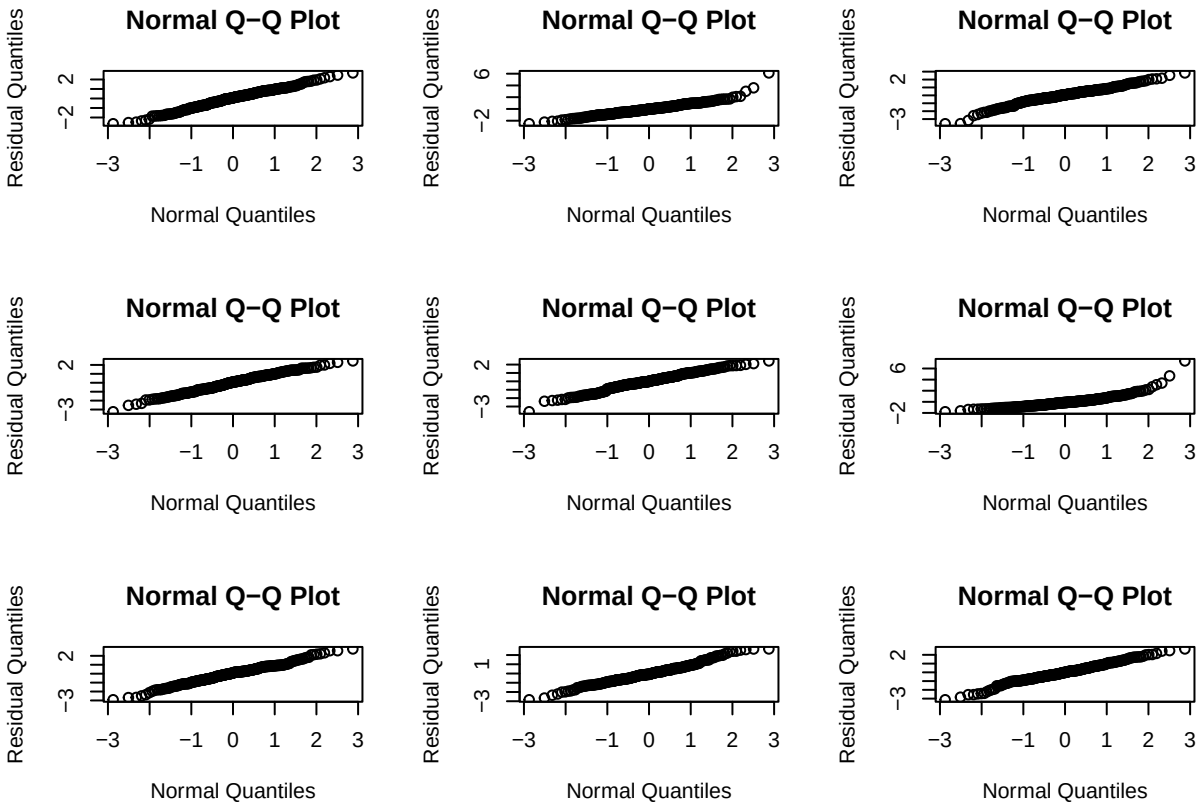


CN

```
rmat2 <- apply(Z2,2,function(o){return(o/sd(o))})
par(mfrow=c(3,3))
for (i in 1:9){
  qqnorm(rmat2[,i],xlab = "Normal Quantiles", ylab = "Residual Quantiles")
}
```




```
par(mfrow=c(3,3))
for (i in 10:18){
  qqnorm(rmat2[,i],xlab = "Normal Quantiles", ylab = "Residual Quantiles")
}
```




```r
par(mfrow=c(3,3))
for (i in 19:21){
  qqnorm(rmat2[,i])
}
```