Antonin Rosa

Hendrik Wellmann

Practical 4 Population Substructure

Question 1

```r
library(MASS)
data <- read.table("Chr21.dat", header = FALSE, sep = " ")
num_individuals <- nrow(data) - 1
cat("Number of individuals:", num_individuals, "\n")
```

```
## Number of individuals: 203
```

```r
variant_cols_indices <- which(grepl("^rs", data[1, ], ignore.case = TRUE))
num_variant <- length(variant_cols_indices)
cat("Number of variants:", num_variant, "\n")
```

```
## Number of variants: 138000
```

```r
missing_percentage <- mean(is.na(data)) * 100
cat("Percentage of missing data:", missing_percentage, "%\n")
```

```
## Percentage of missing data: 0 %
```

Question 2

```r
genotype_data <- data[-1, variant_cols_indices]
genotype_data[is.na(genotype_data)] <- 0
manhattan_dist_matrix <- dist(genotype_data, method = "manhattan")
submatrix <- as.matrix(manhattan_dist_matrix)[1:5, 1:5]
print(submatrix)
```

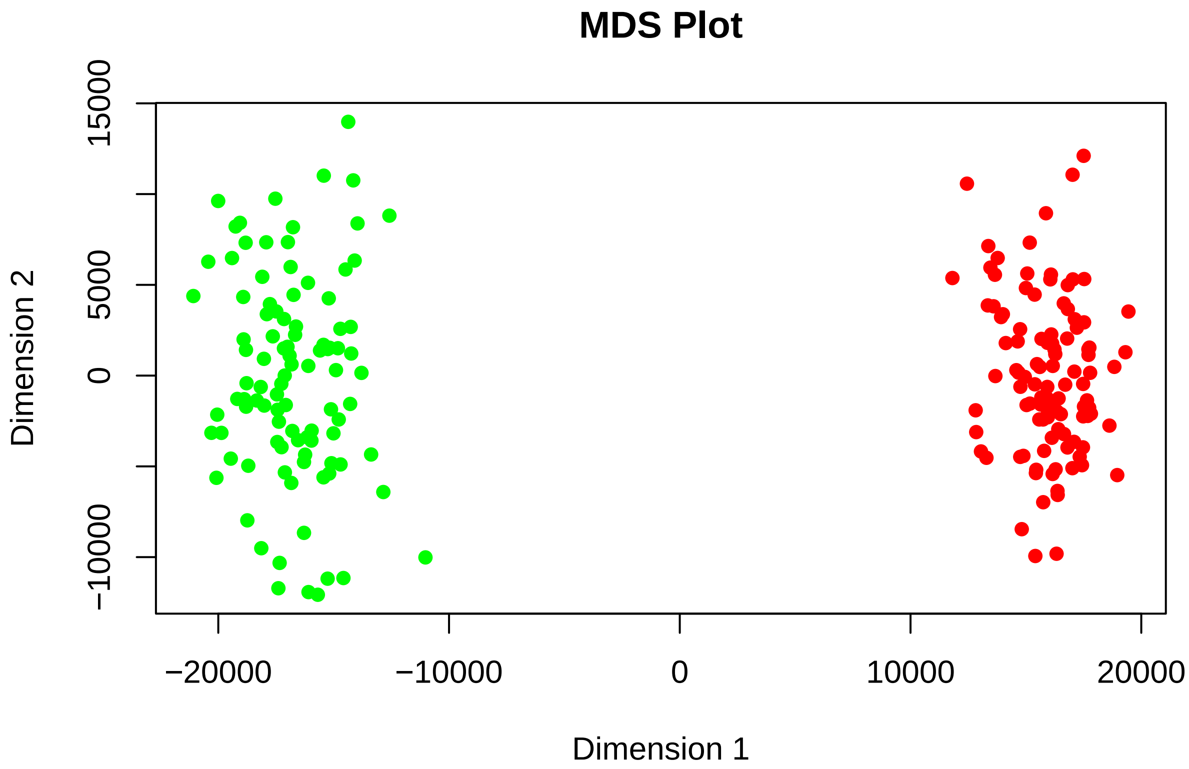
```
##       2     3     4     5     6
## 2     0 53463 54971 58127 53754
## 3 53463     0 55336 55966 55663
## 4 54971 55336     0 54776 55643
## 5 58127 55966 54776     0 59003
## 6 53754 55663 55643 59003     0
```

Question 3

There is an inverse relationship between Manhattan distance and allele sharing. That is, as Manhattan distance increases (indicating greater dissimilarity), allele sharing tends to decrease (indicating lower genetic similarity), and vice versa.

Question 4

```r
mds_result <- cmdscale(manhattan_dist_matrix, k = 2, eig=TRUE)
colors <- ifelse(mds_result$points[ , 1] > 0, "red", "green")
plot(mds_result$points[, 1], mds_result$points[, 2], col = colors, pch = 16, main = "MDS Plot", xlab =
```



```
population1 <- sum(mds_result$points[, 1] > 0)
population2 <- sum(mds_result$points[, 1] < 0)
cat("Population 1 : ", population1, "\n")
```

```
## Population 1 :  104
```

```
cat("Population 2 : ", population2, "\n")
```

```
## Population 2 :  99
```

Two distinct subpopulations are observed.

Question 5

```
ev <- mds_result$eig
gof <- mds_result$GOF
print(gof)
```

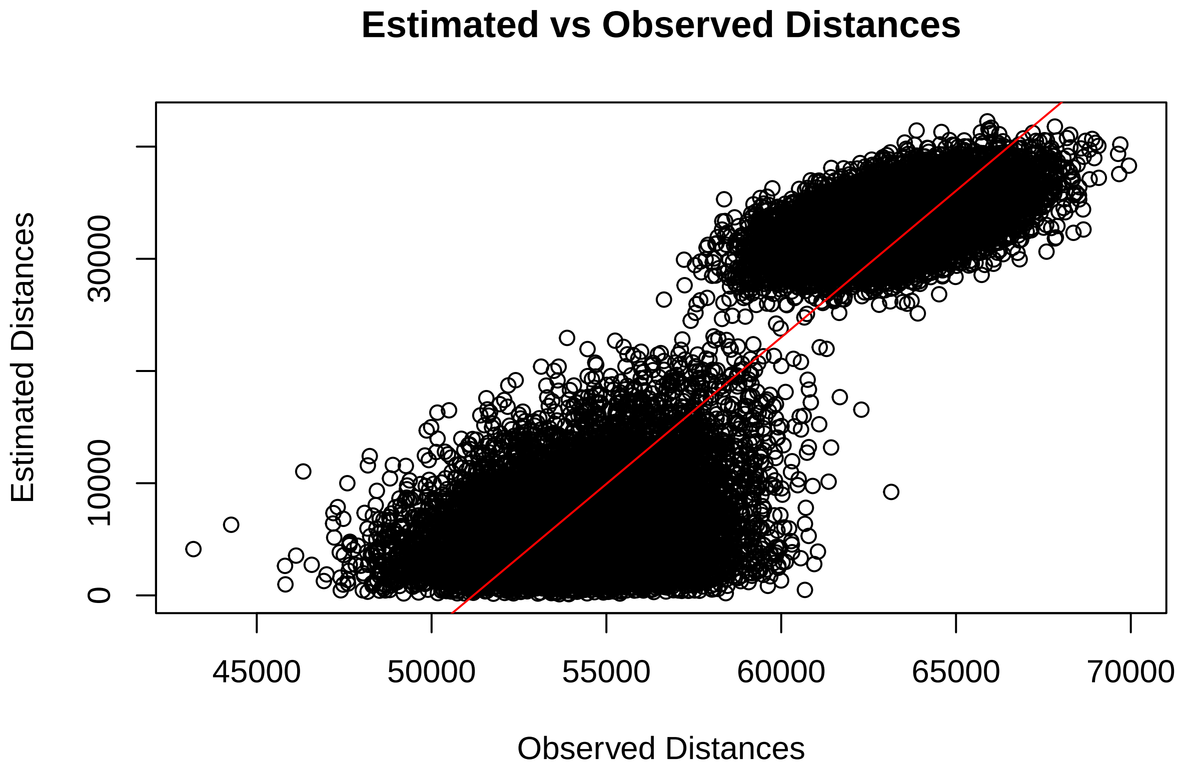
```
## [1] 0.1703840 0.1703865
```

The goodness-of-fit (GOF) of the two-dimensional approximation to your distance matrix is a measure of how well the reduced-dimensional representation preserves the pairwise distances present in the original data. In the context of Multidimensional Scaling (MDS), GOF is often calculated using stress, a criterion that reflects the dissimilarity between the original distances and the distances in the reduced space.

The values we got, suggest a relatively good fit.

Question 6

```
mds_result <- cmdscale(manhattan_dist_matrix, k = 2)
observed_distances <- as.vector(manhattan_dist_matrix)
estimated_distances <- as.vector(dist(mds_result))

plot(observed_distances, estimated_distances,
     main = "Estimated vs Observed Distances",
     xlab = "Observed Distances", ylab = "Estimated Distances")

regression_model <- lm(estimated_distances ~ observed_distances)
abline(regression_model, col = "red")
```



```
cat("Coefficient of Determination (R-squared):", summary(regression_model)$r.squared, "\n")
```

```
## Coefficient of Determination (R-squared): 0.8428011
```

We can observe that there is a big difference between the estimated distances and the observed distance.

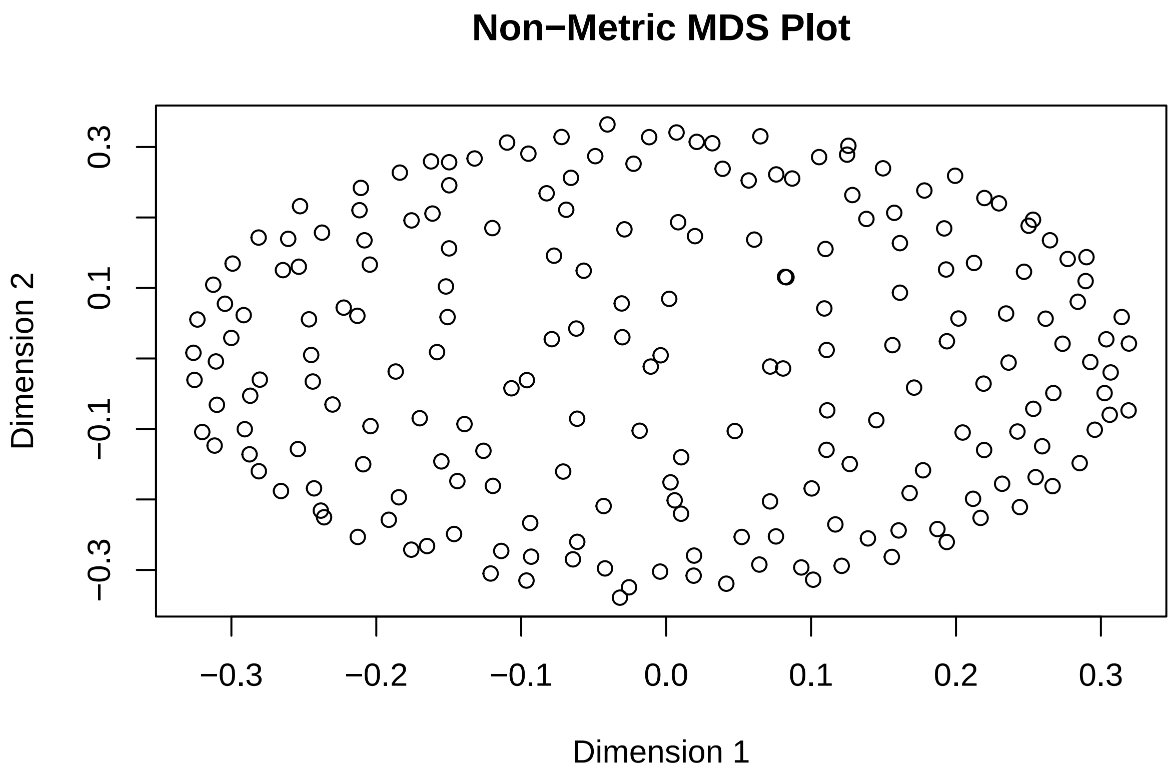
Question 7

```
manhattan_matrix <- as.matrix(manhattan_dist_matrix)
n <- nrow(manhattan_matrix)
m <- ncol(manhattan_matrix)

set.seed(12345)
random_matrix <- matrix(runif(n * m), nrow = n, ncol = m)
```

```
nonmetric_mds_result <- isoMDS(random_matrix, k = 2)
```

```
## initial  value 44.788979
## iter   5 value 41.382387
## final  value 41.302849
## converged
```

```
plot(nonmetric_mds_result$points[, 1], nonmetric_mds_result$points[, 2],
     main = "Non-Metric MDS Plot", xlab = "Dimension 1", ylab = "Dimension 2")
text(nonmetric_mds_result$points[, 1], nonmetric_mds_result$points[, 2], labels = rownames(nonmetric_mds
```



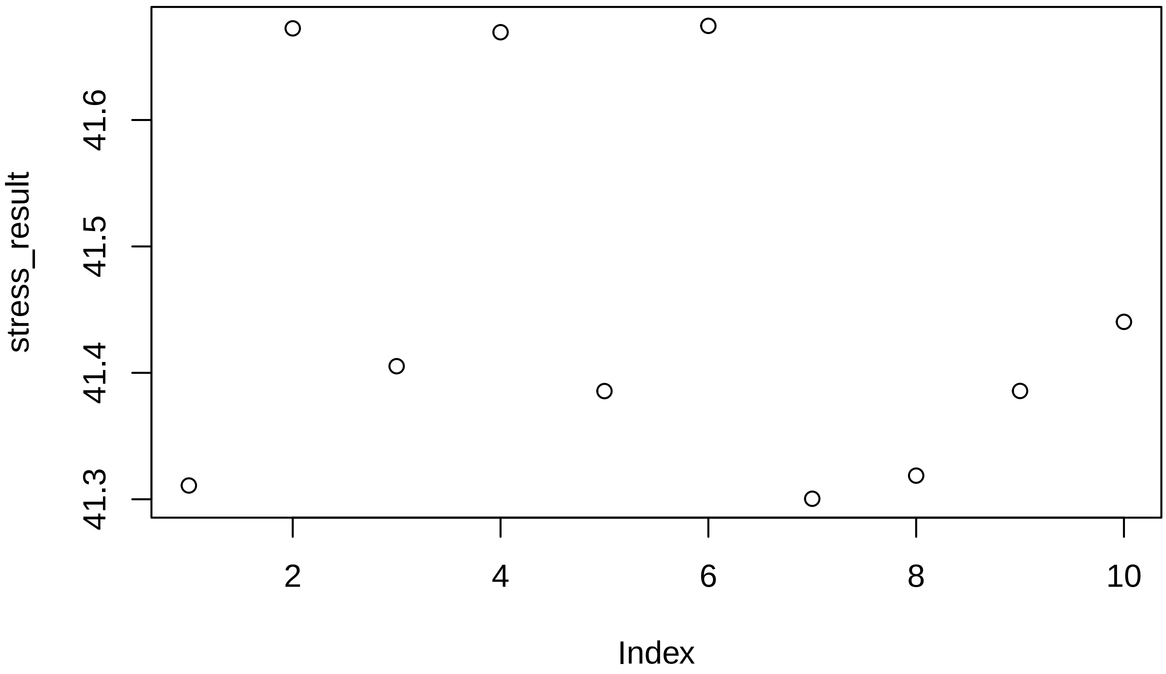
```
nonmetric_mds_result$stress
```

```
## [1] 41.30285
```

It seems the data comes from a same population because we can see only one cluster. However, the stress value shows we should increase the dimension of our MDS.

Question 8

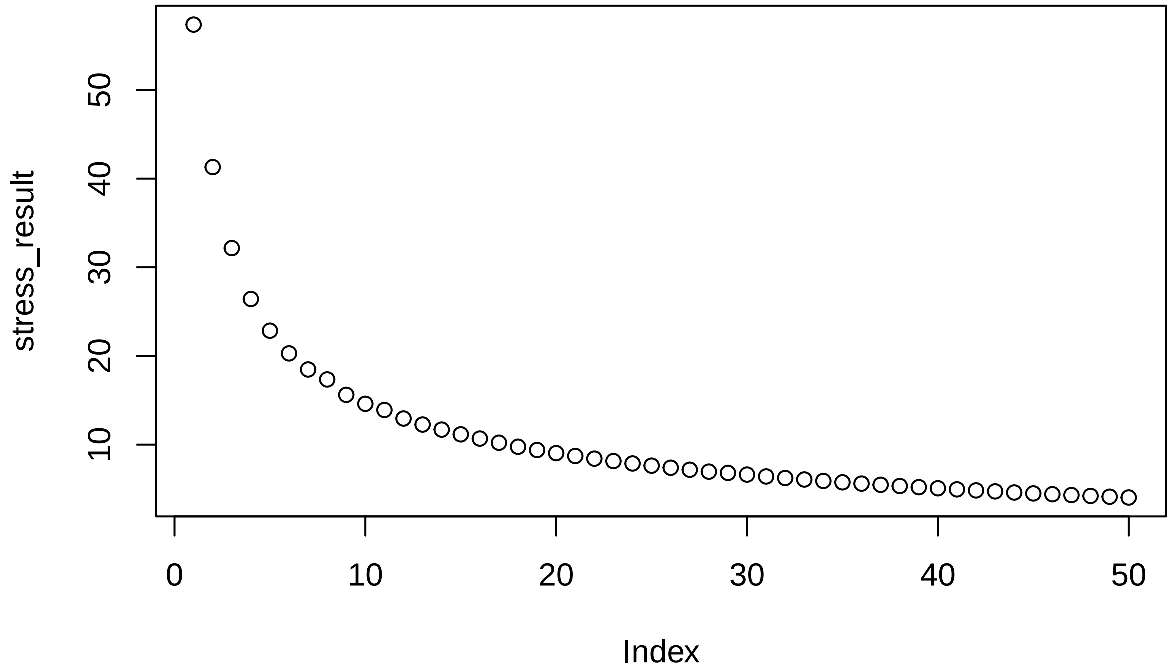
```
num_runs <- 10
stress_result <- c()
for (i in 1:num_runs) {
  set.seed(i)
  random_matrix <- matrix(runif(n * m), nrow = n, ncol = m)
  mds_result <- isoMDS(random_matrix, k = 2, trace=FALSE)
  stress_result <- c(stress_result, mds_result$stress)
}

plot(stress_result)
```



We observe that in all the situation the stress value is very similar each time.

Question 9

```
set.seed(12345)
stress_result <- c()
random_matrix <- matrix(runif(n * m), nrow = n, ncol = m)
for (i in 1:50) {
  set.seed(i)
  mds_result <- isoMDS(random_matrix, k = i, trace=FALSE)

  stress_result <- c(stress_result, mds_result$stress)
}

plot(stress_result)
```



We need 18 dimensions.

Question 10

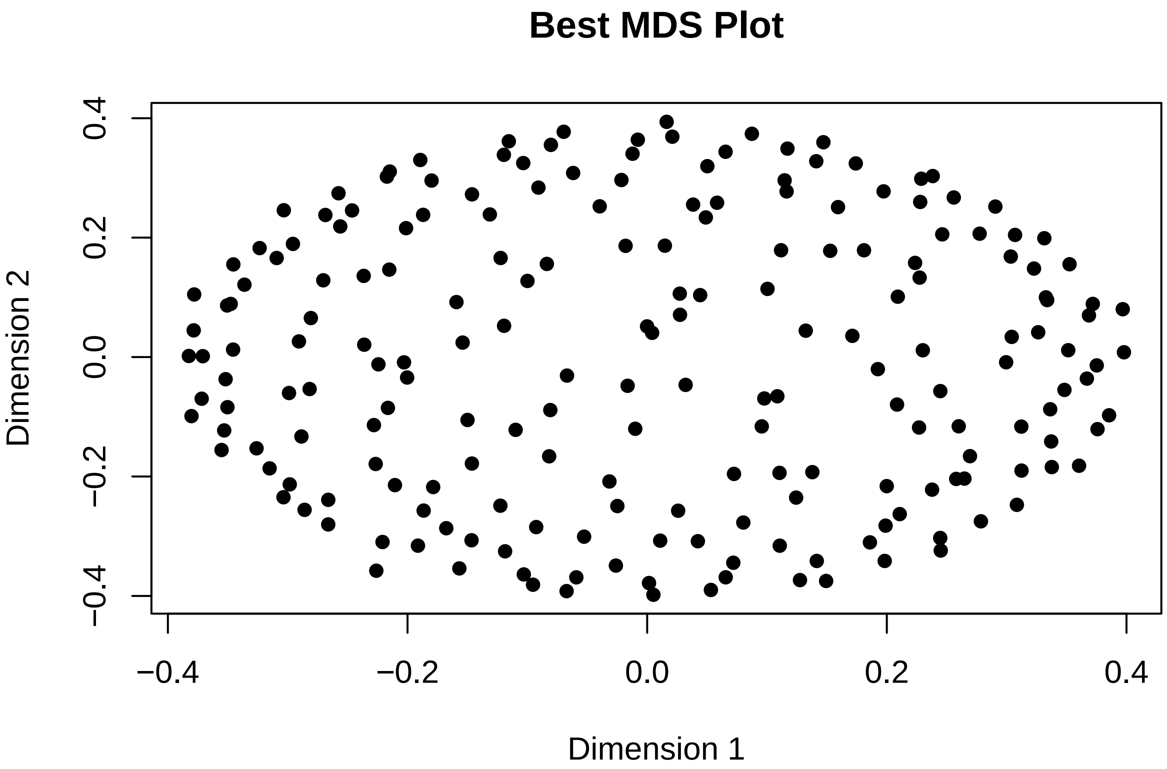
```
num_runs <- 100
stress_result <- c()
point_result <- vector("list", num_runs)
for (i in 1:num_runs) {
  set.seed(i)
  random_matrix <- matrix(runif(n * m), nrow = n, ncol = m)
  mds_result <- isoMDS(random_matrix, k = 2, trace=FALSE)

  stress_result <- c(stress_result, mds_result$stress)
  point_result[[i]] <- mds_result$points
}
cat("best:", min(stress_result))
```

```
## best: 41.02987
```

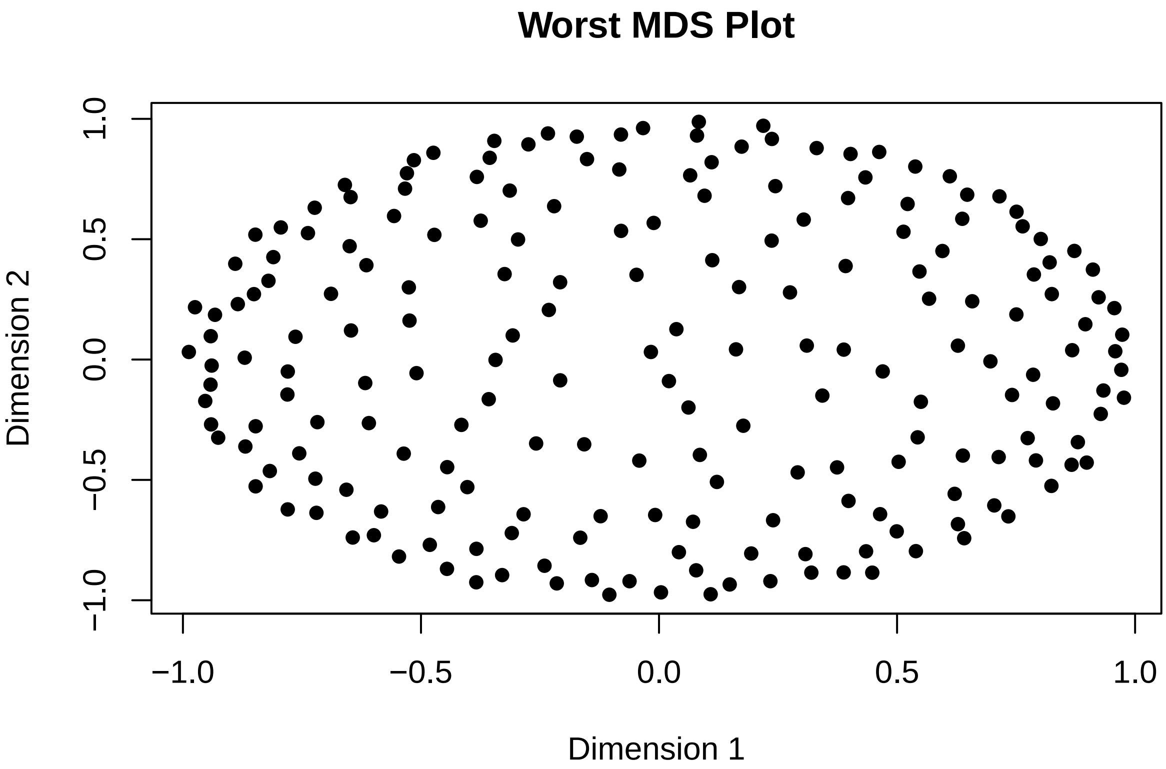
```
best_result = point_result[which.min(stress_result)][[1]]
cat("worst:", max(stress_result))
```

```
## worst: 41.67448
```

```
worst_result = point_result[which.max(stress_result)][[1]]
plot(best_result[, 1], best_result[, 2], pch = 16, main = "Best MDS Plot", xlab = "Dimension 1", ylab =
```



```
plot(worst_result[, 1], worst_result[, 2], pch = 16, main = "Worst MDS Plot", xlab = "Dimension 1", ylab
```



You can see that the data for the best results is closer together than the data for the worst result (range -0.4 to 0.4 compared to range -1 to 1).

Question 11

```
set.seed(which.min(stress_result))
random_matrix <- matrix(runif(n * m), nrow = n, ncol = m)
metric_result <- cmdscale(random_matrix, k = 2)
correlation_matrix <- cor(best_result, metric_result)
correlation_matrix
```

```
##            [,1]        [,2]
## [1,] 0.89282660 -0.01120042
## [2,] 0.01409546  0.87728440
```

The strong positive correlations on the diagonal indicate that the overall patterns of variation in the first and second dimensions are similar between the metric and non-metric MDS solutions. However, the low correlation with the off-diagonal element suggests that there may be differences in the specific patterns of variation captured by these dimensions.

The high positive correlations on the diagonal suggest that the two-dimensional solutions are capturing similar overall structures, even though there might be differences in the specific details of how points are arranged in the reduced space.