Ma 322: Biostatistics
Homework Assignment 11

Prof. Wickerhauser

Read Chapter 16, "Working with Multivariate Data," pages 288–318 of our text.

1. Load the NCI data set from the class web site and write `R` codes to do the following:

   (a) Print the list of cancers that appear 7 or more times in the 64 rows.

   (b) Count the total number of rows of data from cancers appearing 7 or more times.

   (c) Print the column names that contain gene expression data with the top 5 variances.

   (d) Plot the classification tree for the rows in part a and the columns in part c.

   (e) Print the misclassification rate for the tree in part d.

   Solution: See the `R` commands and output in `hw11R.txt`.

   (a) See the `R` commands in `hw11R.txt`, similar to those in `r-eg-27.txt`, that find the cancers that appear at least 7 times.

   (b) See the `R` commands in `hw11R.txt`, similar to those in `r-eg-27.txt`, that extract the rows with cancers that appear at least 7 times.

   (c) See the `R` commands in `hw11R.txt`, similar to those in `r-eg-27.txt`, that extract the columns with the top 5 variances.

   (d) See the `R` commands in `hw11R.txt`, similar to those in `r-eg-27.txt`, that produced the plot in Figure HW11,Ex.1d below.

   (e) Misclassification error rate: $0.3 = 12/40$ □

2. Use the Fisher `iris` data set with the functions in the `cluster` library.

   (a) Use `agnes()` to compute, and then plot, the dendrogram for the iris data decomposed by agglomerative hierarchical clustering.

   (b) Use `diana()` to compute, and then plot, the dendrogram for the iris data decomposed by divisive hierarchical clustering.

   (c) Use `kmeans()` to find 3 clusters in the iris data and determine the number of misclassifications of the 150 plants.

   (d) Use `pam()` to find 3 clusters in the iris data and determine the number of misclassifications of the 150 plants.

   Solution: See the file `hw11R.txt` for the `R` commands and their output.

   (a) See Figure HW11,Ex.2(a) below.

   (b) See Figure HW11,Ex.2(b) below.

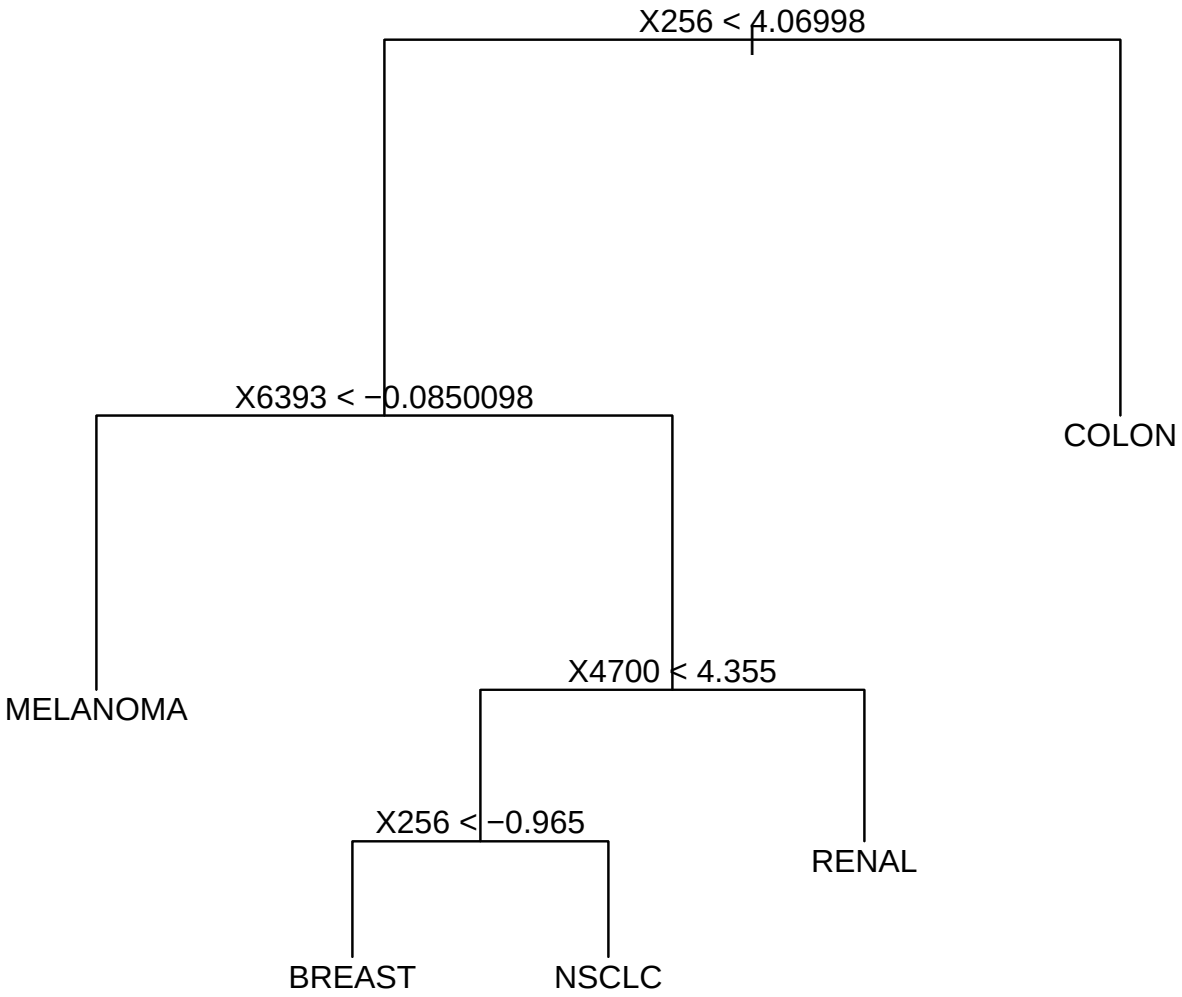


Figure 1: HW11,Ex.1(d): Classification tree for NCI cancer data: 40 cancers with at least 7 replications; top 5 genes by expression variance.

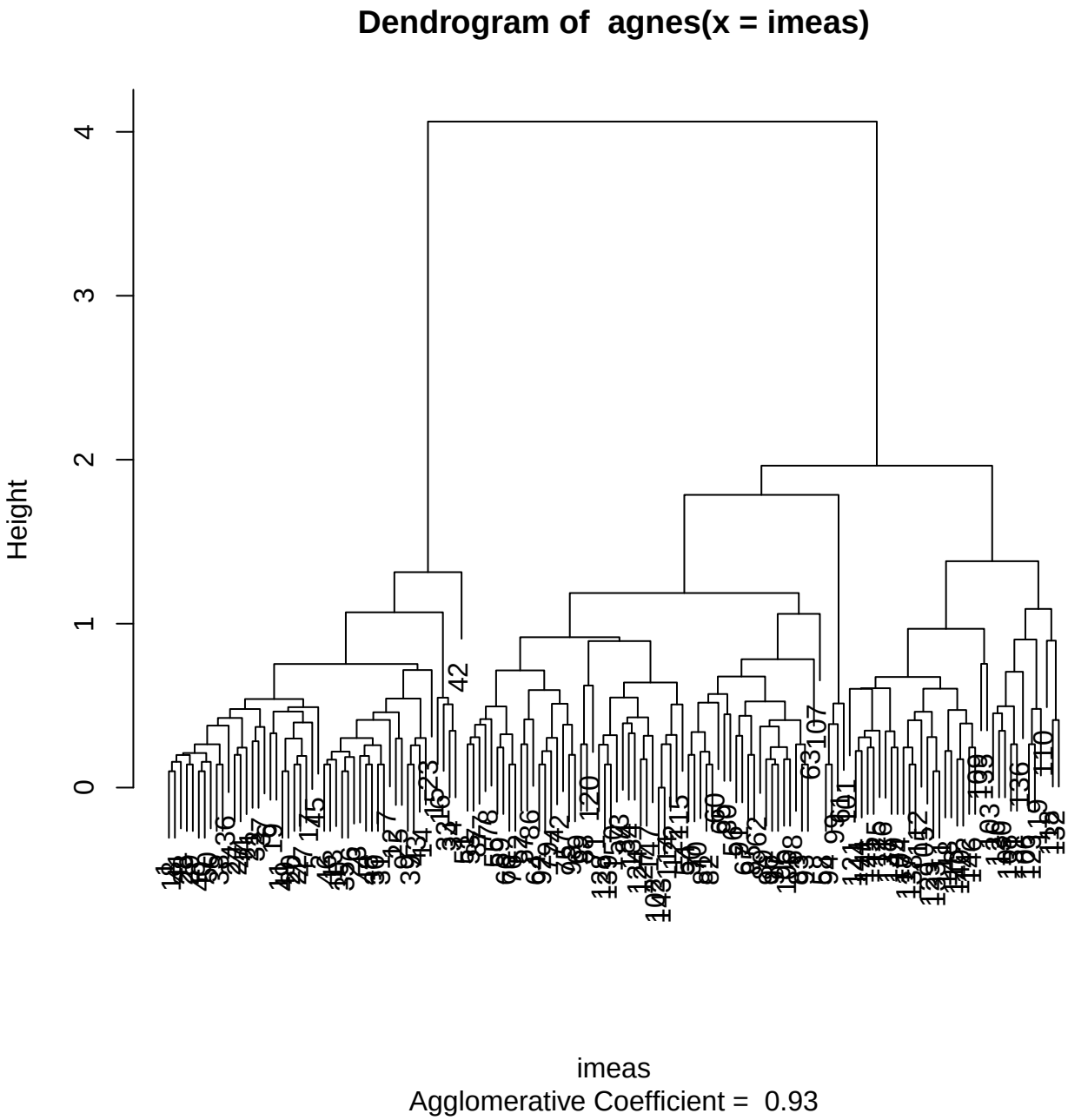


Figure 2: HW11,Ex.2(a): Dendrogram for agglomerative hierarchical clustering of the Fisher iris data.

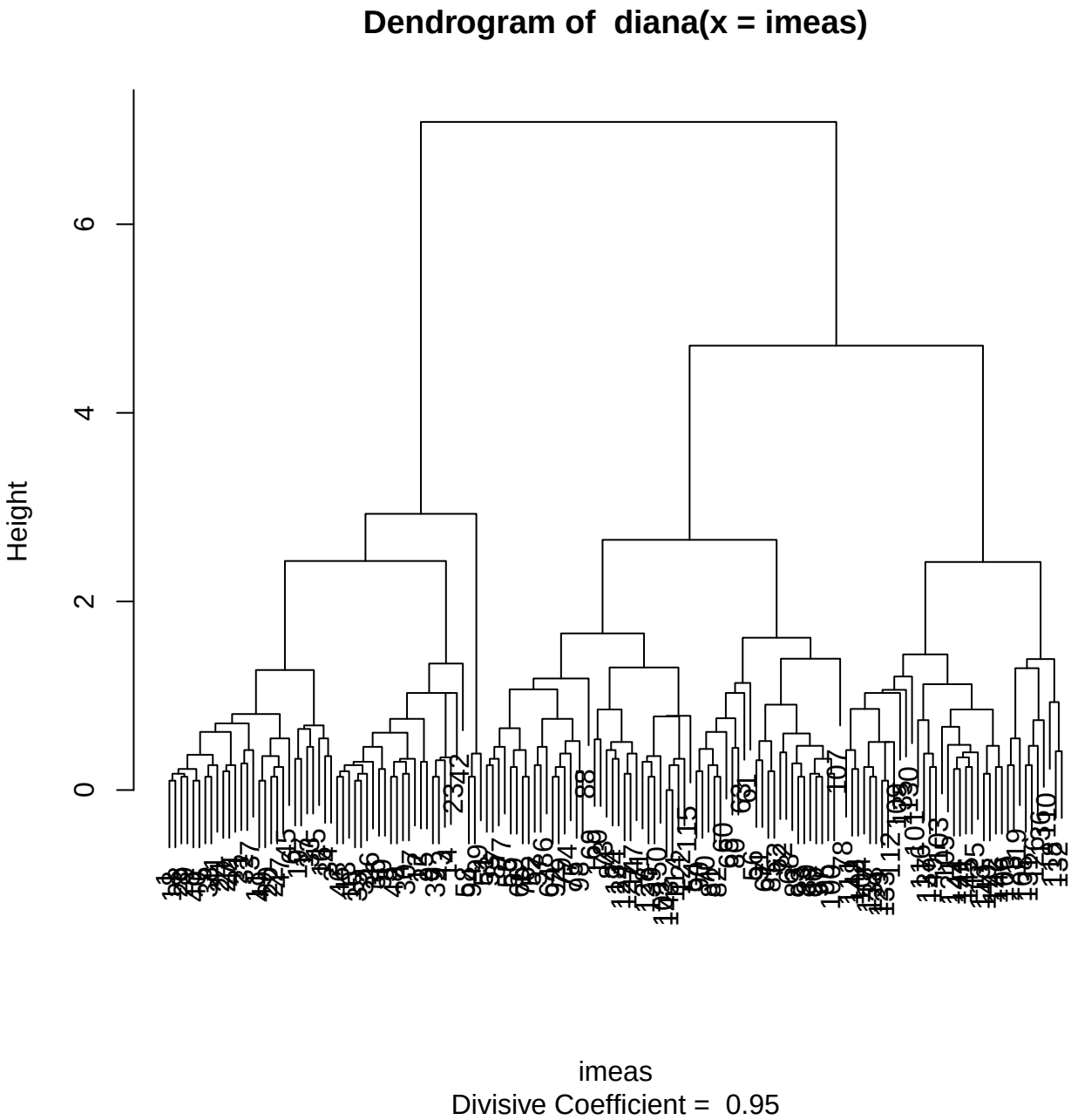


Figure 3: HW11,Ex.2(b): Dendrogram for divisive hierarchical clustering of the Fisher iris data.

(c) With `kmeans()`, there were 71 misclassifications, out of 150.

(d) With `pam()`, there were 16 misclassifications, out of 150. □

3. Use the data set of 57 cancers with at least 3 repetitions, with the top 12 expressed genes, produced by the commands in `http://www.math.wustl.edu/~victor/classes/ma322/r-eg-27.txt`,

   This data may also be found in `http://www.math.wustl.edu/~victor/classes/ma322/nci57x13.R` and, after copying that file into the folder being used by your `R` session, can be read into data frame `nci` with the command

   ```
   load("nci57x13.R")
   ```

   Install and load the `vegan` library into `R` and use `isomap()` for the following:

   (a) Find the Euclidean distances between samples. Using $k = 2$ nearest neighbors, apply multidimensional scaling to these Euclidean distances and plot the result.

   (b) Find the Manhattan distances between samples. Using $k = 2$ nearest neighbors, apply multidimensional scaling to these Manhattan distances and plot the result.

   Solution: (a) See Figure HW11,Ex.3(a) below.

   (b) See Figure HW11,Ex.3(b) below. □

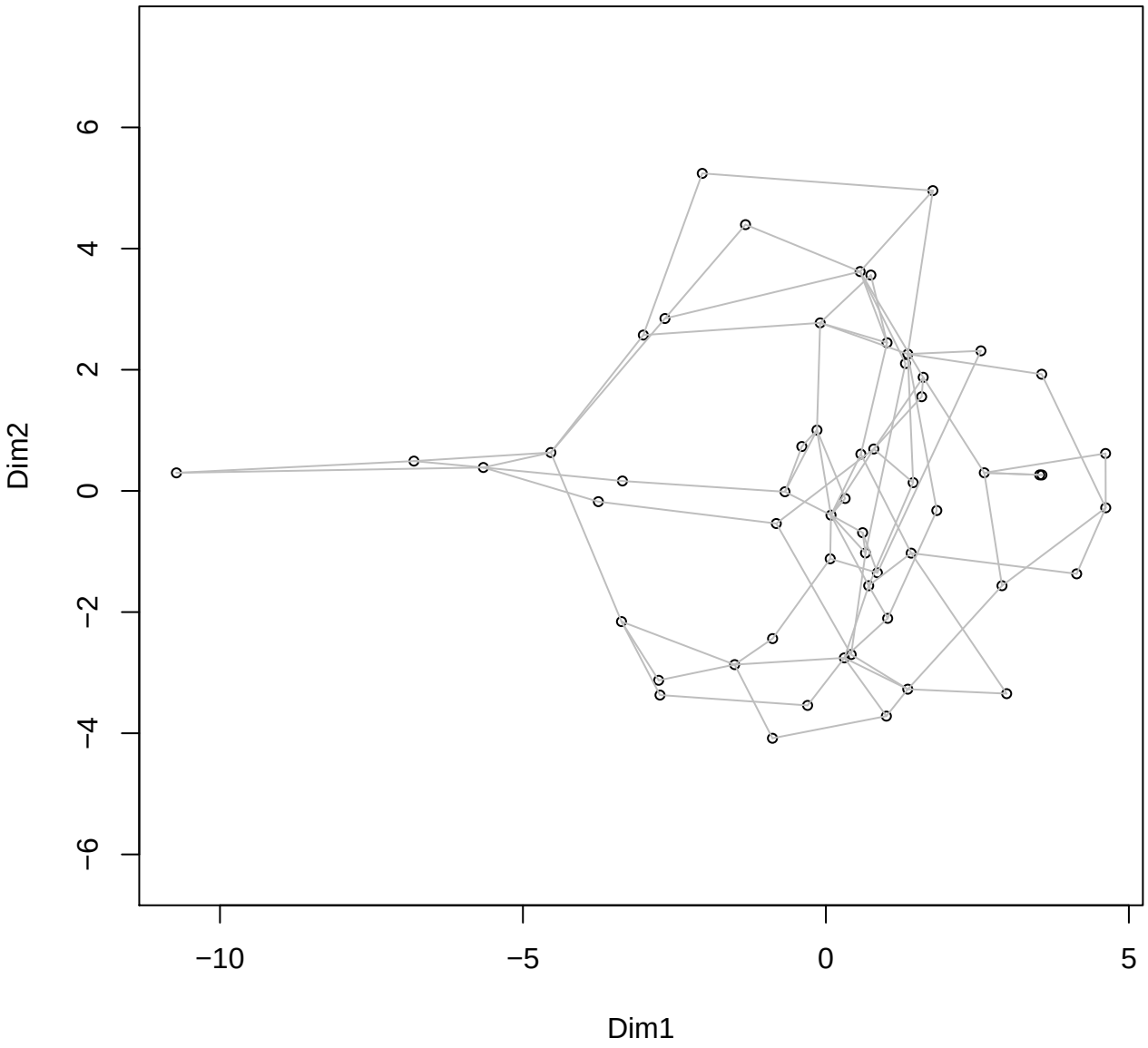


Figure 4: HW11,Ex.3(a): Isomap plot of 57 cancer data by top 12 expressed genes, starting with 2 nearest neighbors by Euclidean distance.

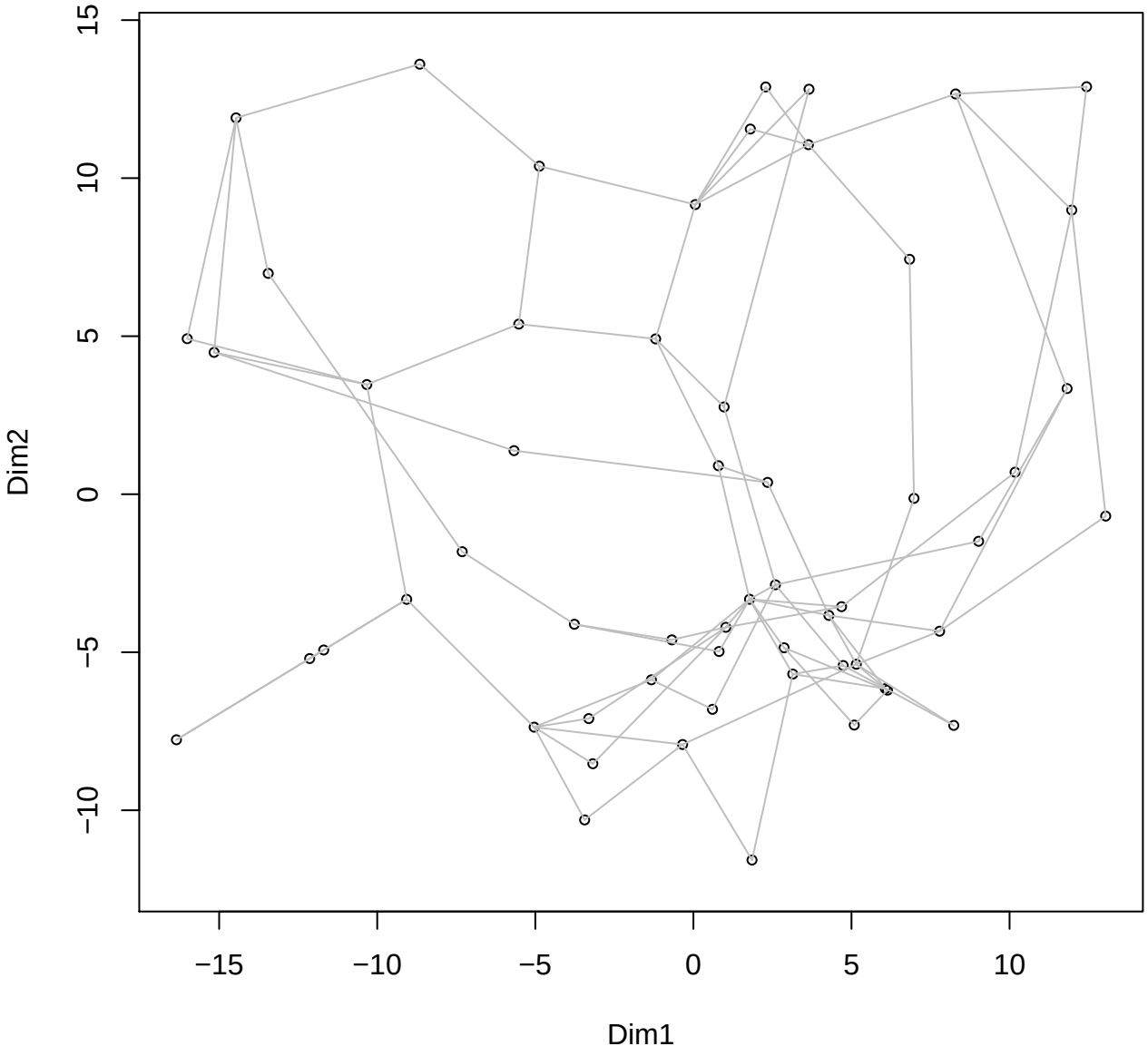


Figure 5: HW11,Ex.3(b): Isomap plot of 57 cancer data by top 12 expressed genes, starting with 2 nearest neighbors by Manhattan distance.