# Supplementary Material for "Graphical Representation and Modeling of Metagenomic Reads"

Deepak Nag Ayyala, Shili Lin

## 1 Simulation Model I

The following section contains graphical representations of one random realization of data generated using Model I. We generated OTU counts for 300 samples equally divided into 10 clusters. For each sample, the counts are generated from 300 OTUs. To distinguish the samples from different clusters, the OTUs are equally partitioned in such a way that observations within a cluster have the same unique set of OTUs that have non-zero counts. Therefore, the models are expected to predict complete separation between the clusters. The non-zero counts are modeled as

$$X_{ij} \sim Poi\left(\lambda_{ij}\right), \qquad \lambda_{ij} \sim NB(10, 1),$$

where "Poi" is the Poisson model with mean $\lambda_{ij}$ and "NB" is the negative binomial model with mean 10 and dispersion parameter 1.

Quantities describing the differences in clustering of the different models are presented

in Table S1. The table includes the estimated number of clusters and misclassification error as defined in Equation (3) of the paper. The misclassification errors presented in the table indicate that the MDS models outperform their PCoA counterparts. The PCoA models are seen to underestimate the number of clusters, as opposed to their MDS counterparts. Graphical representations of both $l_1$-MDS and PCoA models are presented in the following subsections.

Table S1. Measures of modeling accuracy for a dataset randomly generated using Model I.

| Model | Misclassification error | Estimated no. of clusters |
|---|---|---|
| 2D $l_1$-MDS | 6.91% | 10 |
| 3D $l_1$-MDS | 2.37% | 10 |
| 4D $l_1$-MDS | 0% | 10 |
| 2D PCoA | 32.11% | 4 |
| 3D PCoA | 30.10% | 5 |
| 4D PCoA | 4.01% | 8 |

## 1.1 Silhouette plot

The average silhouette plot of all the six models is presented in Figure S1. The optimal number of clusters are selected using the mPAM procedure described in Section 2.4 of the paper. The optimal number of clusters selected for the 2, 3 and 4 dimensional MDS models are all equal to 10. The number of clusters for the three PCoA models are 4, 5 and 8 respectively.

## 1.2 $l_1$-MDS models

Figure S2 presents the two dimensional $l_1$-MDS models showing the estimated and true clusters in the data. Snapshots of the three dimensional models are presented in Figure S3. Interactive graphical models can be accessed by clicking on the links provided in the

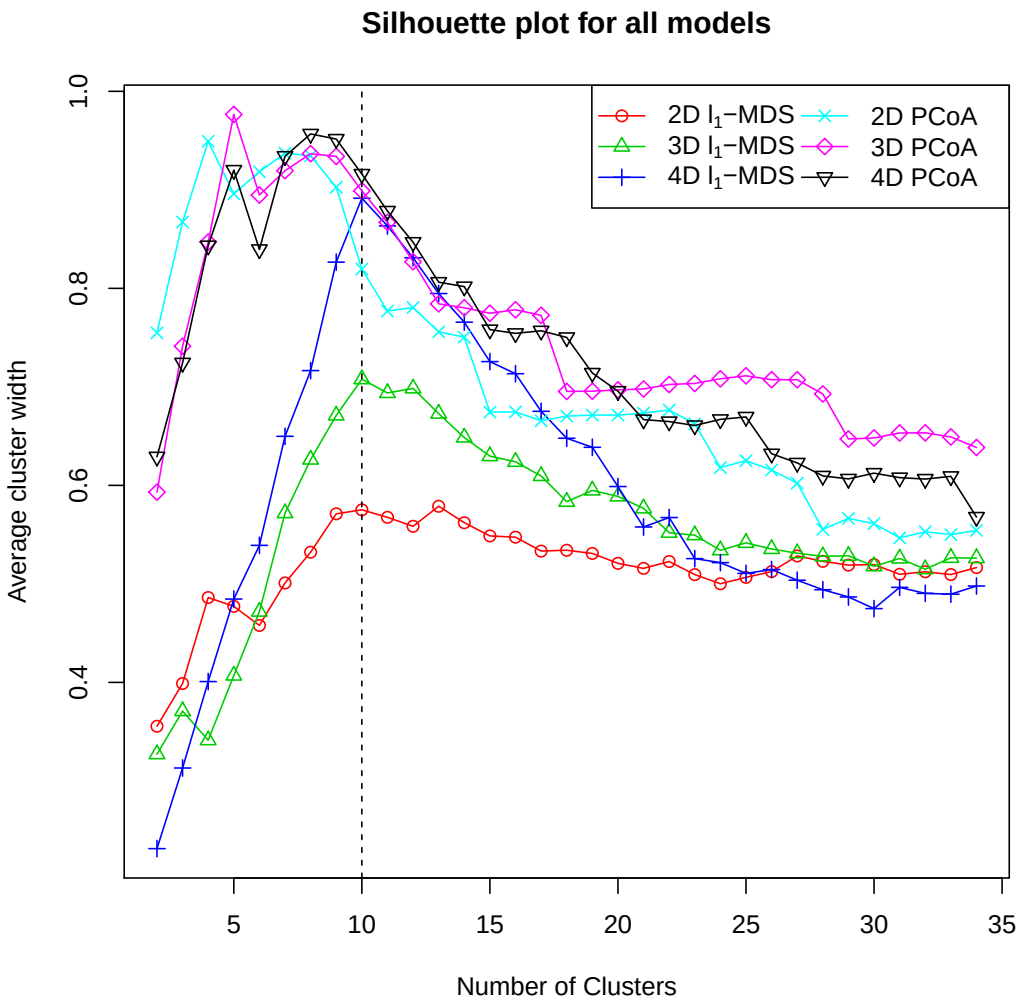



Figure S1: Average silhouette plot of all the six models for the data randomly generated using Model I.

figure. Since four dimensional models are impossible to visualize, we can construct four graphs ignoring one axis at a time. Figures S4 and S5 show snapshots of the four 4D models with executable links to interactive graphs. The four plots showing the 4D model emphasizes the importance of examining all four orientations to visualize all the clusters. For example, the XYZ-axes plot shows clusters colored yellow and pink to be distinct yet the shape of the pink cluster is better visualized in the YZW-axes plot.

## 1.3 PCoA models

In this subsection, PCoA models of all three dimensions are presented. As seen from the silhouette plot(Figure S1), the optimal number of clusters for 2, 3 and 4 dimensions are estimated as 4, 5 and 8 respectively. The two dimensional models are presented in Figure S6, which show the estimated and true clusters. It can be seen that seven clusters in the true model are being estimated as a single color(colored red in the estimated model). Snapshots of the 3D model, presented in Figure S7, display similar behavior of combining clusters in the estimated model. Snapshots and links to interactive graphs for the four dimensional models are presented in Figures S8 and S9.

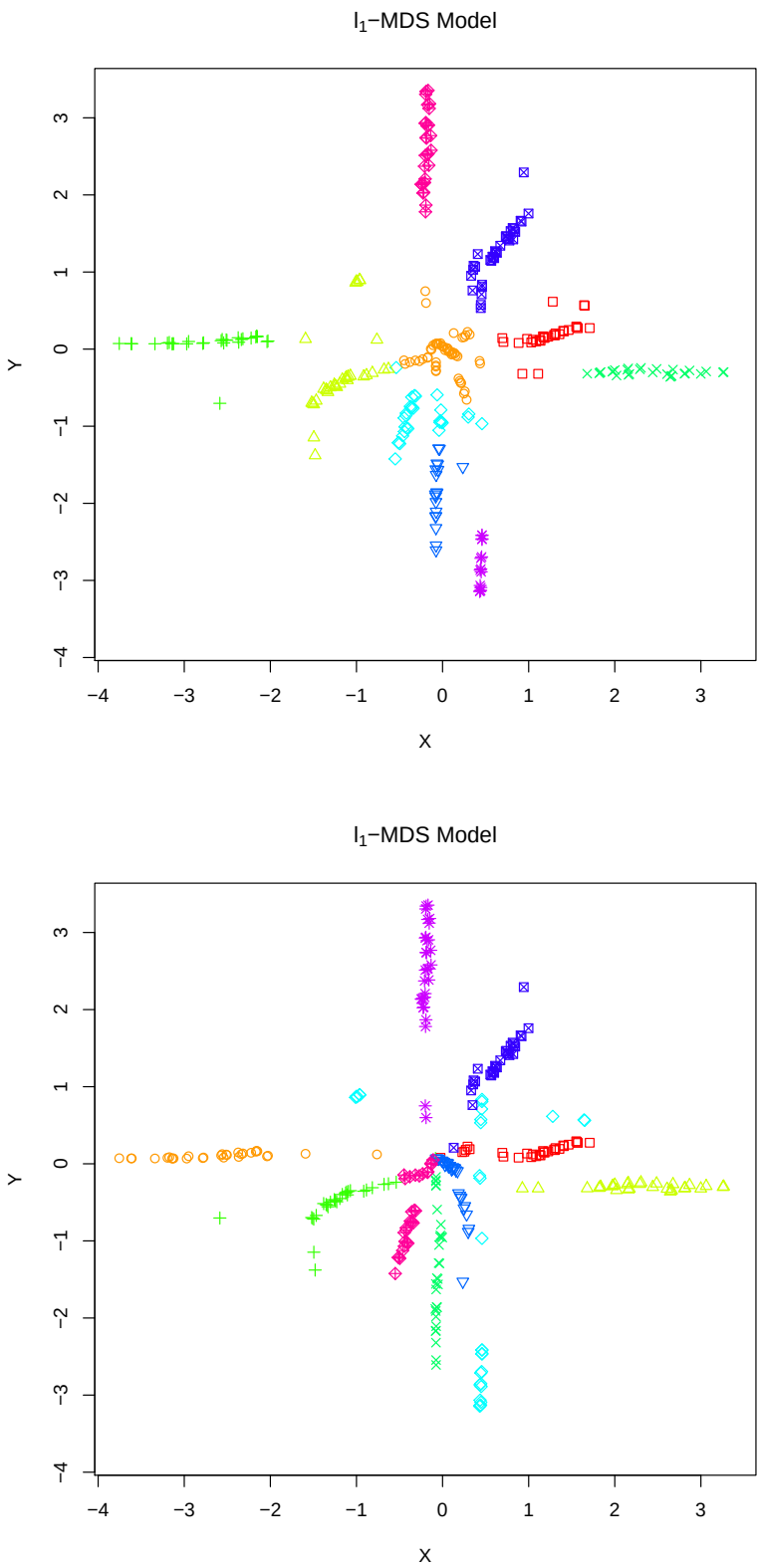
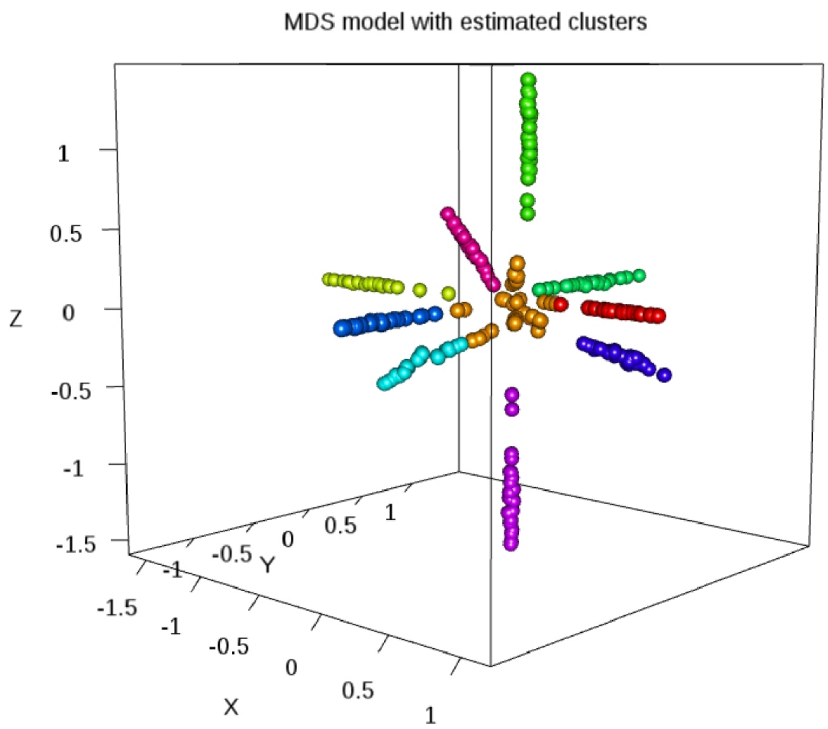



Figure S2: Two dimensional $l_1$-MDS models displaying the estimated and true clusters for the generated data. The estimated and true number of clusters are both equal to 10.

Estimated clusters




True clusters

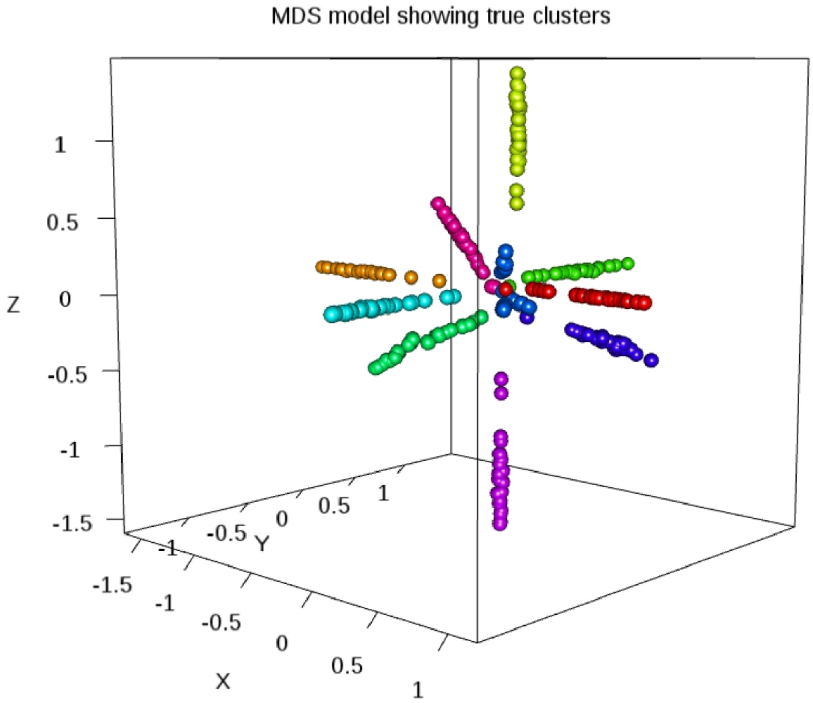



Figure S3: Three dimensional $l_1$-MDS models displaying the estimated and true clusters for the generated data. Interactive models can be opened by clicking the links provided above the graphs.

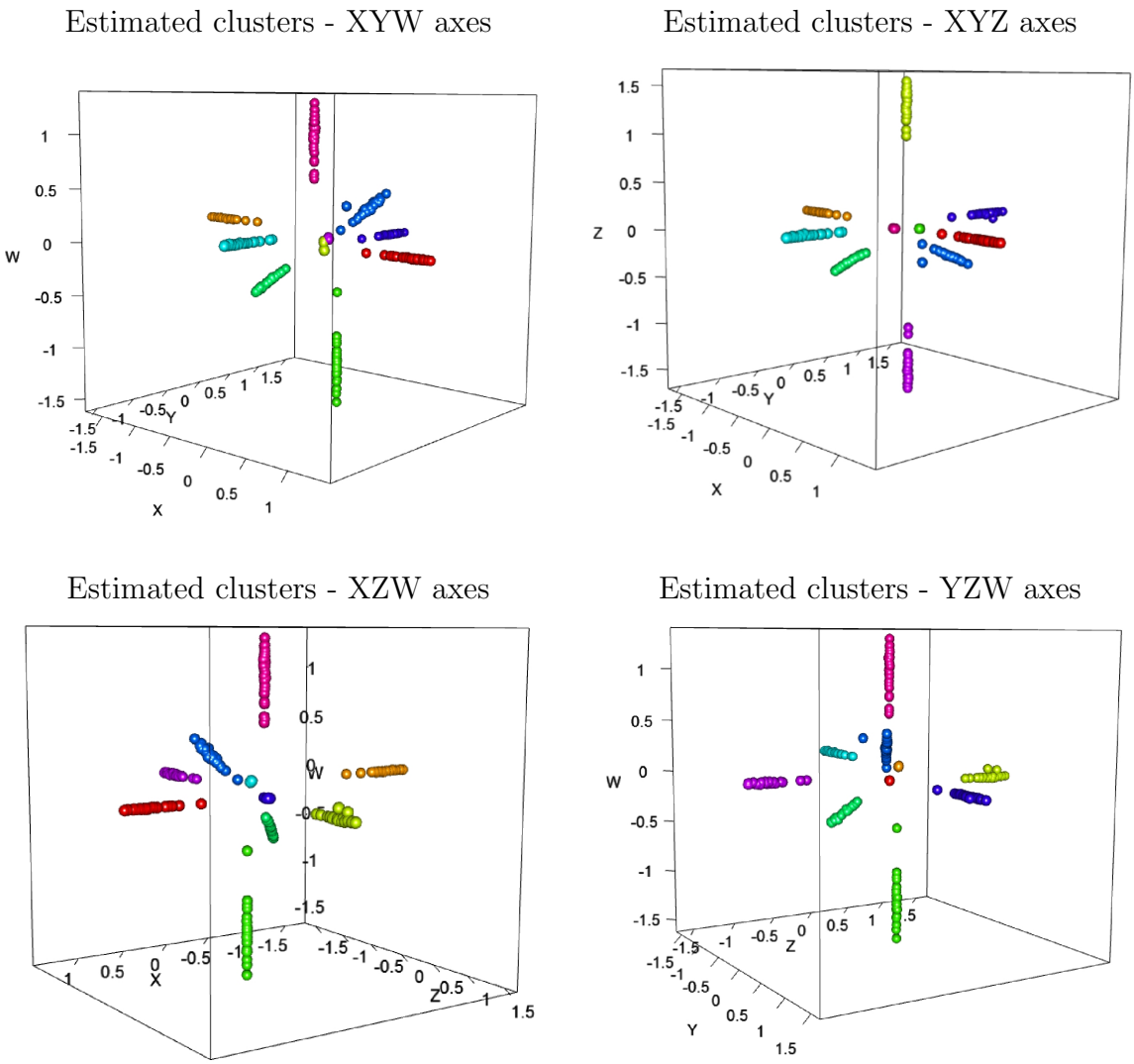



Figure S4: Snapshots of the four dimensional $l_1$-MDS models showing the estimated clusters constructed by removing one axis at a time.

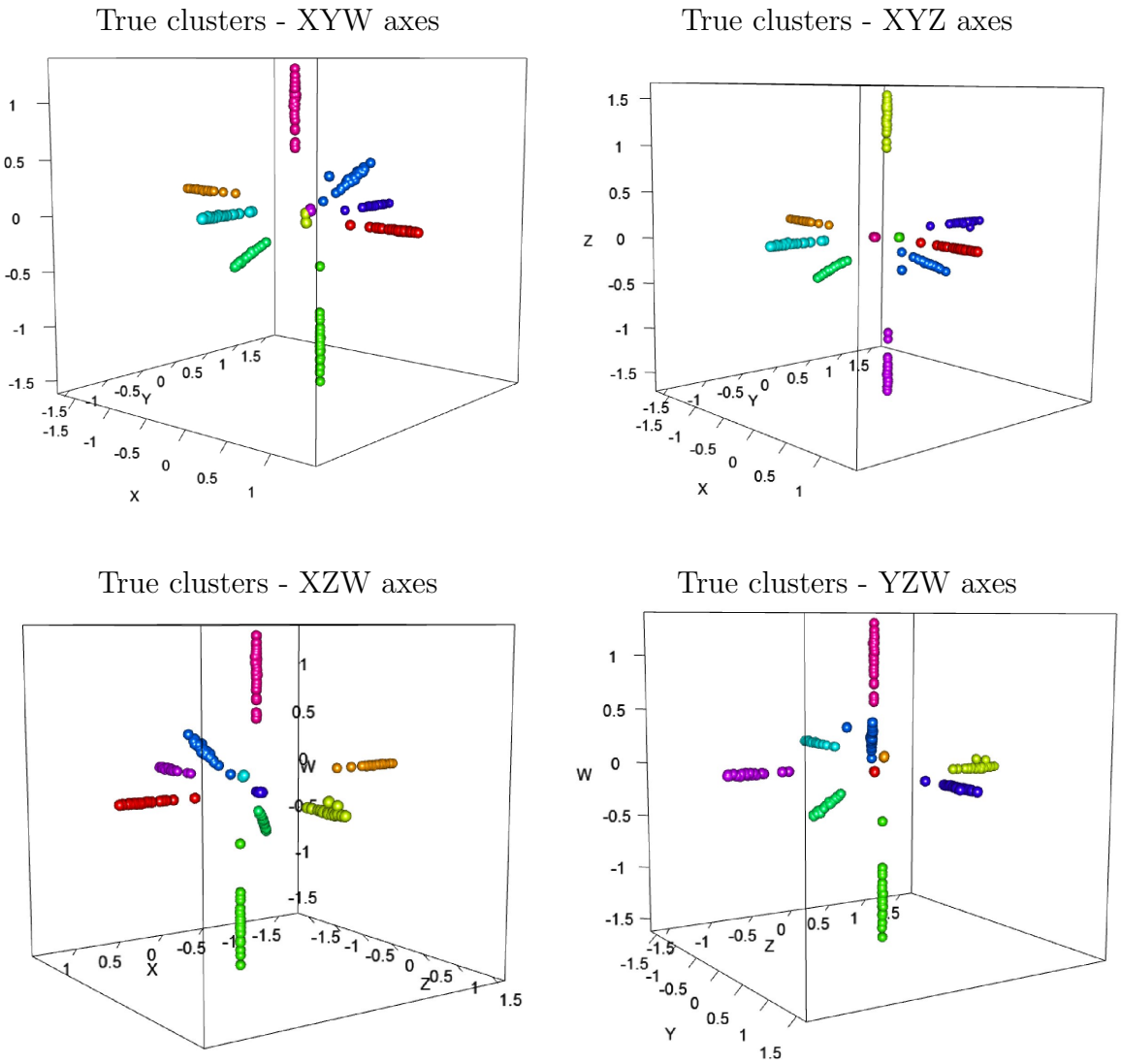
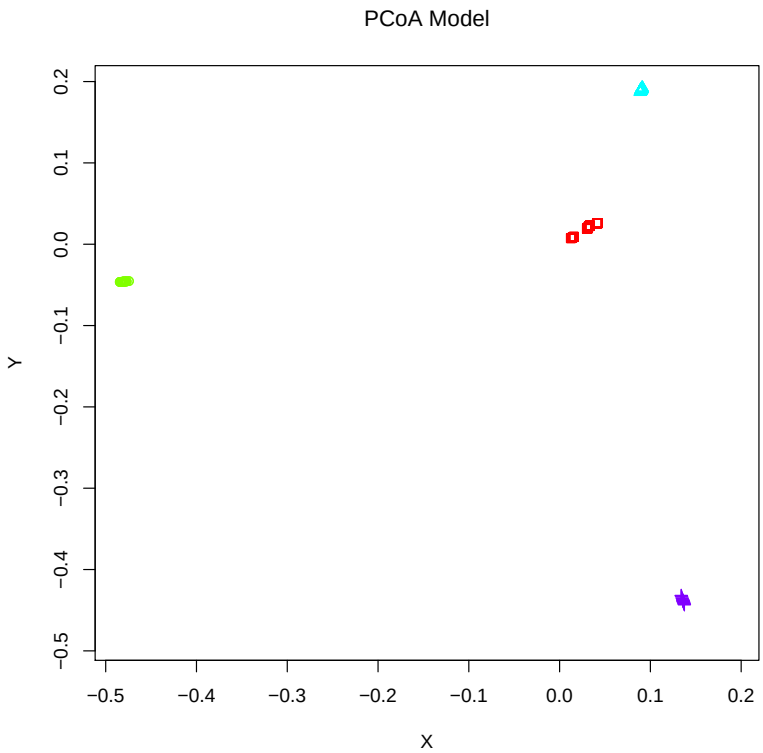



Figure S5: Snapshots of the four dimensional $l_1$-MDS models showing the true clusters constructed by removing one axis at a time.



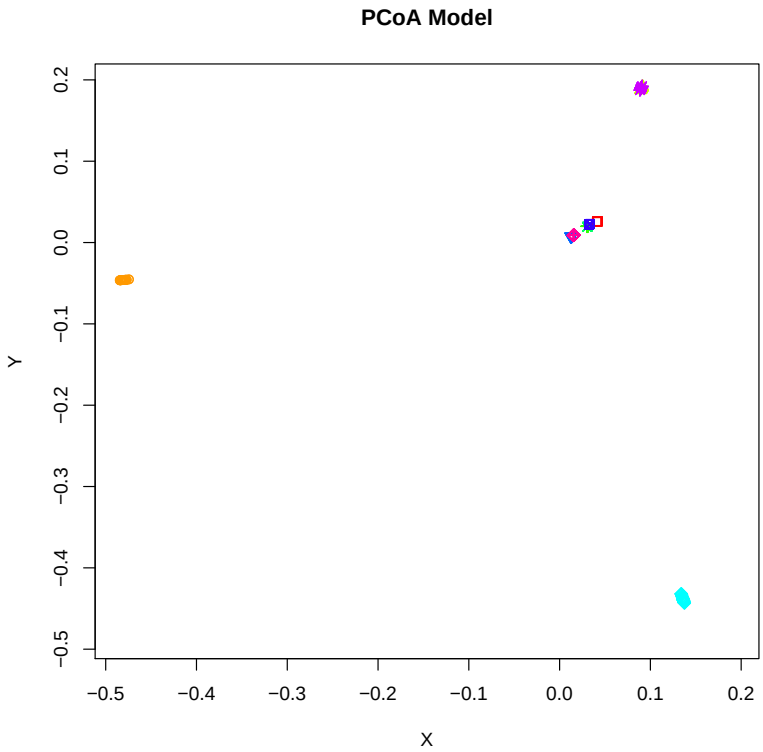
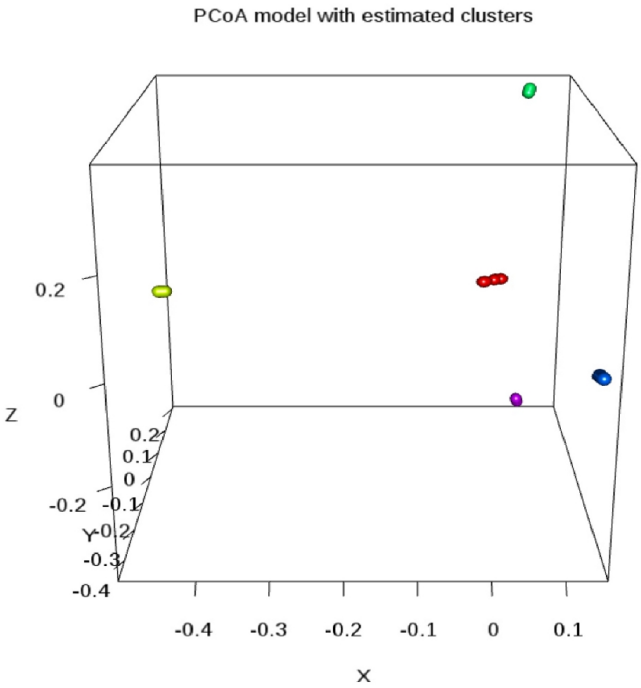




Figure S6: Two dimensional MDS models displaying the estimated and true clusters for the generated data. The estimated number of clusters is 4 and the true number is equal to 10.

Estimated clusters




True clusters

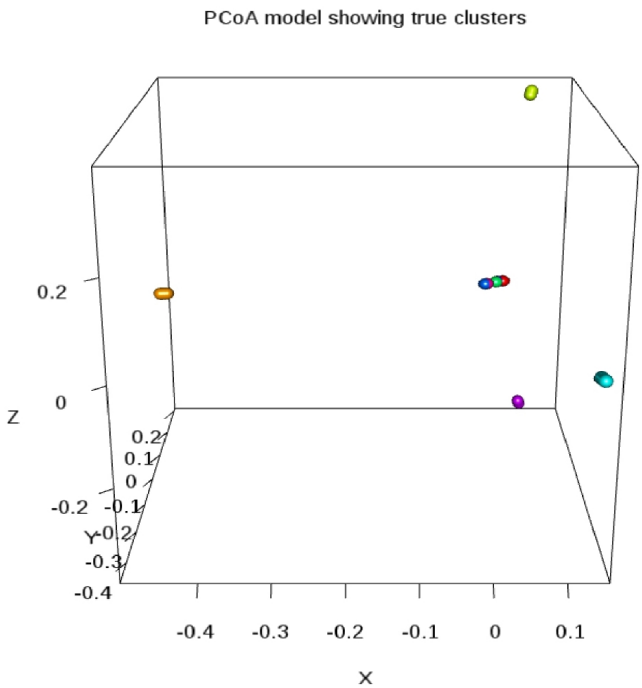



Figure S7: Snapshopts of three dimensional PCoA models displaying the estimated and true clusters. Interactive models can be opened by clicking on the links provided above the graph.

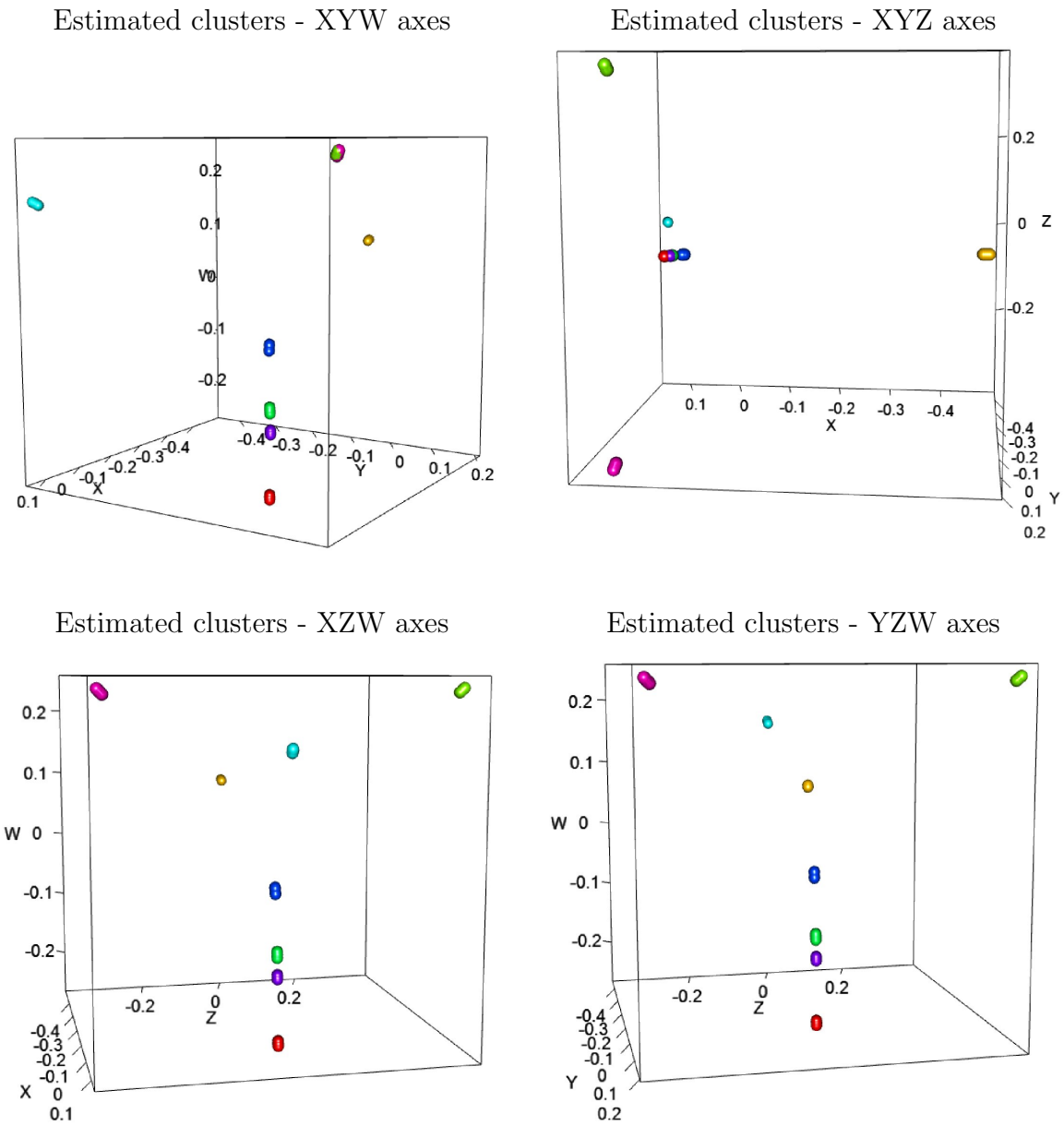



Figure S8: Snapshots of the four dimensional PCoA models showing the estimated clusters constructed by removing one axis at a time.

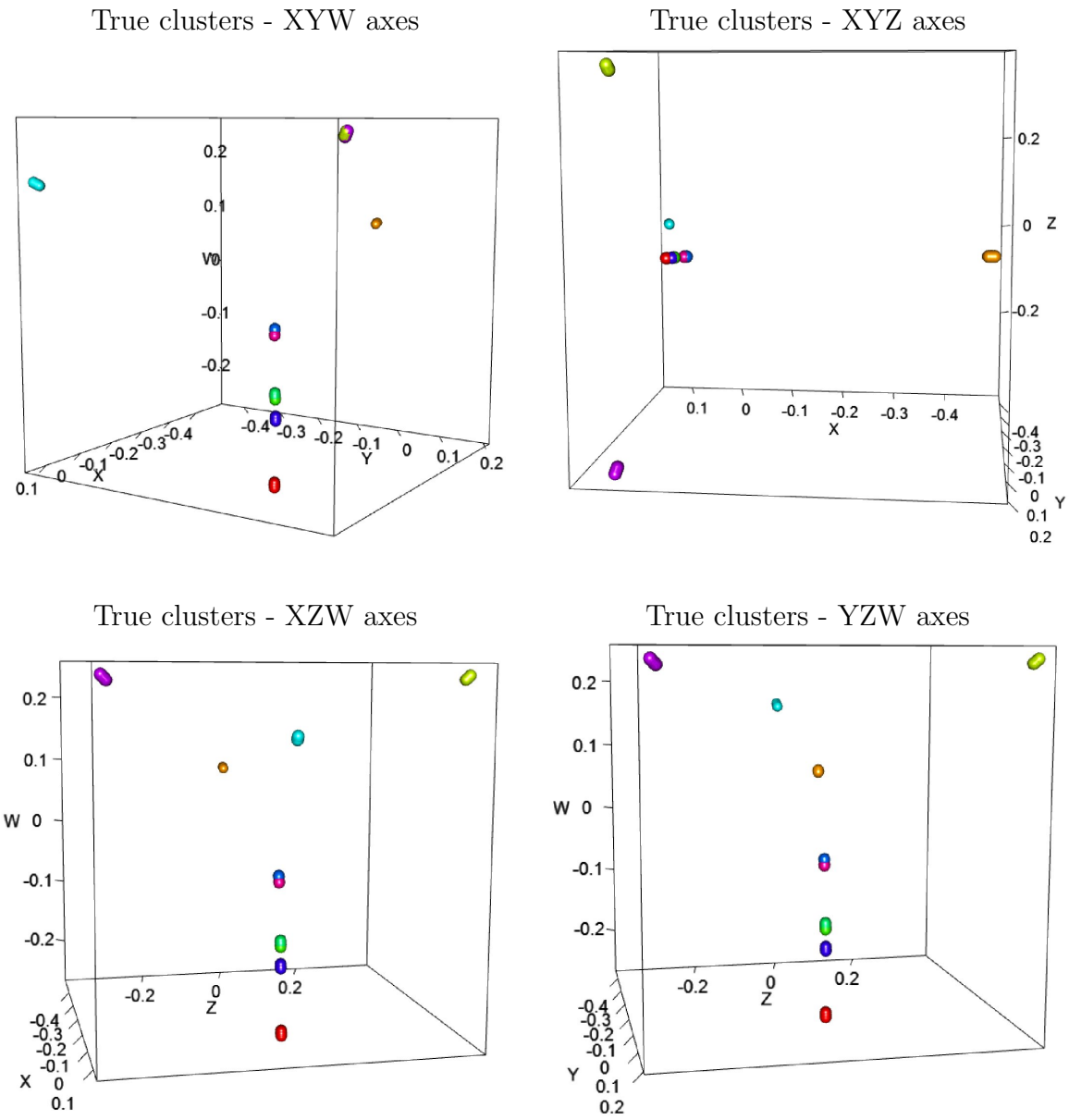



Figure S9: Snapshots of the four dimensional PCoA models showing the true clusters constructed by removing one axis at a time.

# 2 Simulation Model II

The following section contains graphical representations of a random realization of data generated using Model II. In this model, 8 clusters are generated consisting of 40 samples each. The number of OTUs is set to be 200. Within each cluster, each OTU was randomly assigned a zero or non-zero count with a probability of 0.3. The counts are generated as

$$X_{ijk} \sim U_{ik} NB(\lambda_{ik}, 1) + (1 - U_{ik})\, NB(10, 0.01),$$

$$\lambda_{ik} \sim NB(100, 1),\ U_{ik} \sim Bernoulli(0.3),$$

for $1 \leq i \leq 8, 1 \leq j \leq 40, 1 \leq k \leq 200$.

That is, for any given cluster, each OTU is selected to have non-zero count with probability 0.3 and held fixed for all samples within the cluster. The non-zero counts are sampled to have means coming from a negative binomial distribution, and the actual counts are in turn generated from a negative binomial distribution. Further, to incorporate sequencing errors for OTUs with zero counts, the zero counts are modeled as a zero-inflated negative binomial variable, given by the second term in the expression for $X_{ijk}$. This two-stage procedure for data generation does not guarantee a distinct separation as in Model I since the OTUs with non-zero counts are not fixed.

For Model II, the misclassification errors and estimated number of clusters for all the six models are tabulated in Table S2. As seen from the table, both $l_1$-MDS and PCoA models cluster the samples identical to the true clusters for dimensions greater than 2. For the two dimensional models, $l_1$-MDS models outperform the PCoA models. Graphical representations of both $l_1$-MDS and PCoA models are presented in the following subsections.

Table S2. Measures of modeling accuracy for a dataset randomly generated using Model II.

| Model | Misclassification error | Estimated no.of clusters |
|---|---|---|
| 2D MDS | 0.86% | 9 |
| 3D MDS | 0% | 8 |
| 4D MDS | 0% | 8 |
| 2D PCoA | 9.4% | 6 |
| 3D PCoA | 0% | 8 |
| 4D PCoA | 0% | 8 |

## 2.1 Silhouette plot

The average silhouette plot of all the six models is presented in Figure S10. The optimal number of clusters selected for the 2, 3 and 4 dimensional$l_1$-MDS models are 9, 8 and 8 respectively, whereas the corresponding number for the PCoA models are 6,8 and 8 respectively. The number of clusters used for generating the data is equal to 8.

## 2.2 $l_1$-MDS models

Figure S11 presents the two dimensional $l_1$-MDS models showing the estimated and true clusters in the data. Snapshots of the three dimensional models are presented in Figure S12. Interactive graphical models can be accessed by clicking on the links provided in the figure. Since four dimensional models are impossible to visualize, we can construct four graphs ignoring one axis at a time. Figures S13 and S14 show snapshots of the four 4D models with executable links to interactive graphs. The four plots showing the 4D model emphasizes the importance of examining all four orientations to visualize all the clusters. For example, the XYZ-axes plot shows clusters colored yellow and pink to be distinct yet the shape of the pink cluster is better visualized in the YZW-axes plot.

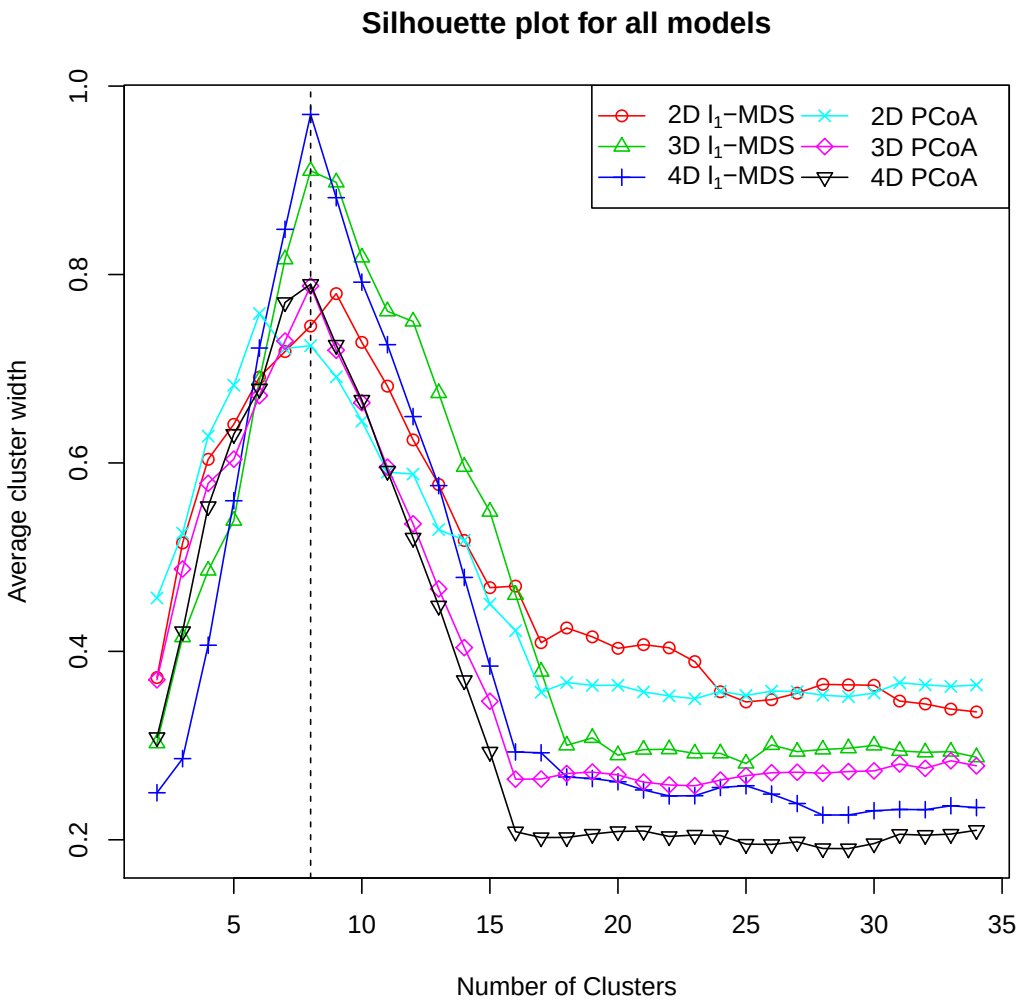



Figure S10: Average silhouette plot of all the six models for data randomly generated using Model II.

## 2.3 PCoA models

In this subsection, PCoA models of all three dimensions are presented. As seen from the silhouette plot(Figure S10), the optimal number of clusters for 2, 3 and 4 dimensions are estimated as 6, 8 and 8 respectively. The two dimensional models are presented in Figure S15, which show the estimated and true clusters. While the models with dimension greater than 2 estimate the clusters correctly, samples within any given cluster have a greater spread compared to the corresponding $l_1$-MDS models. Hence $l_1$-MDS models

seem to be advantageous in cases where the separation between clusters is smaller or when the number of clusters is large.

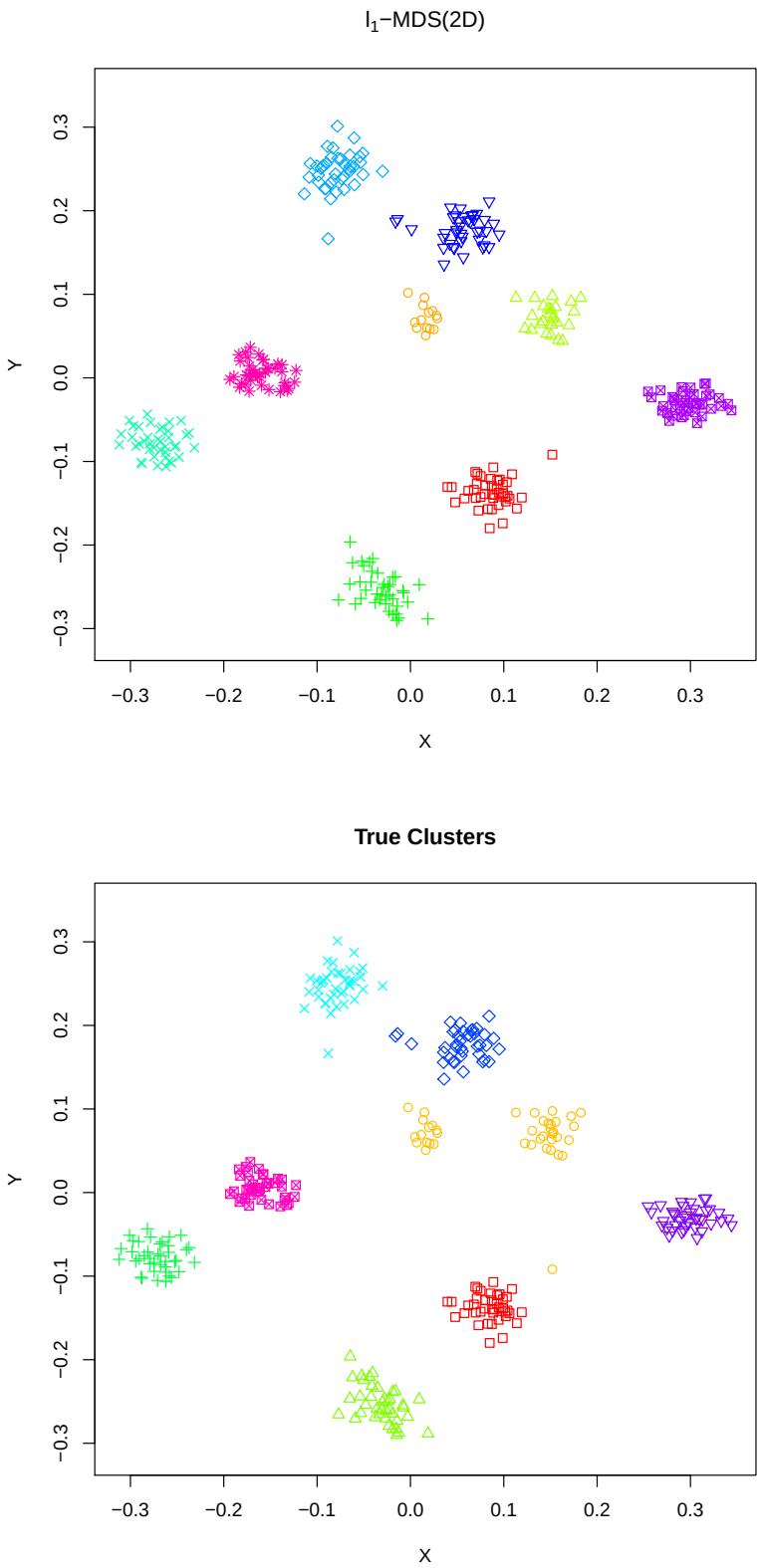
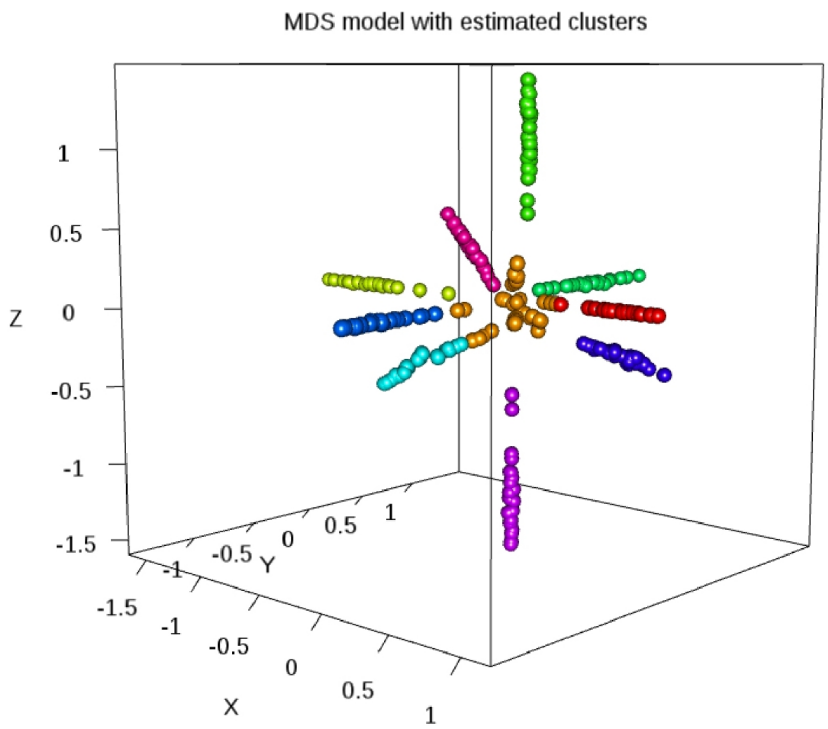



Figure S11: Two dimensional $l_1$-MDS models displaying the estimated and true clusters for the generated data.

Estimated clusters




True clusters

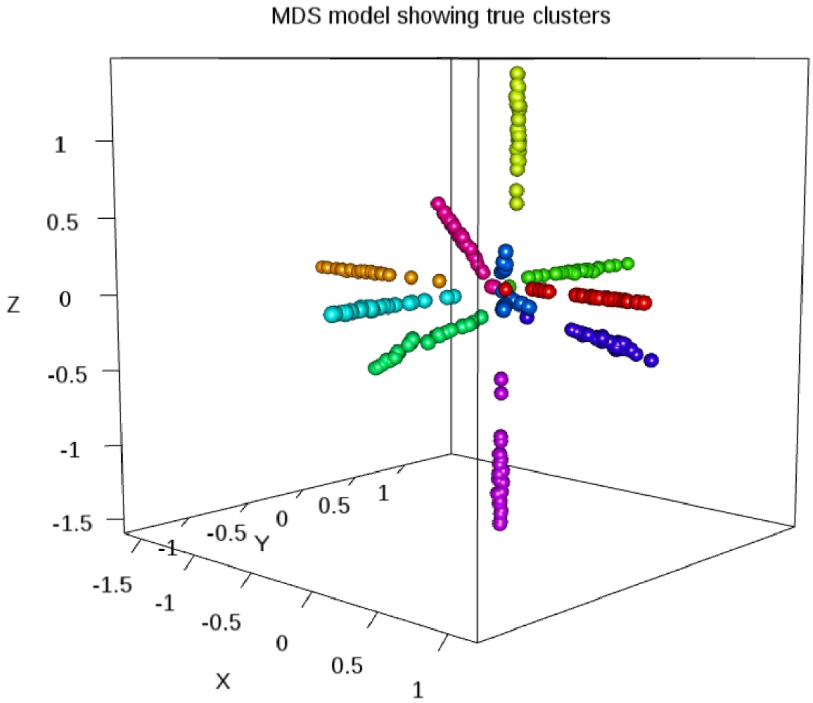



Figure S12: Three dimensional $l_1$-MDS models displaying the estimated and true clusters for the generated data. Interactive models can be opened by clicking the links provided above the graphs.

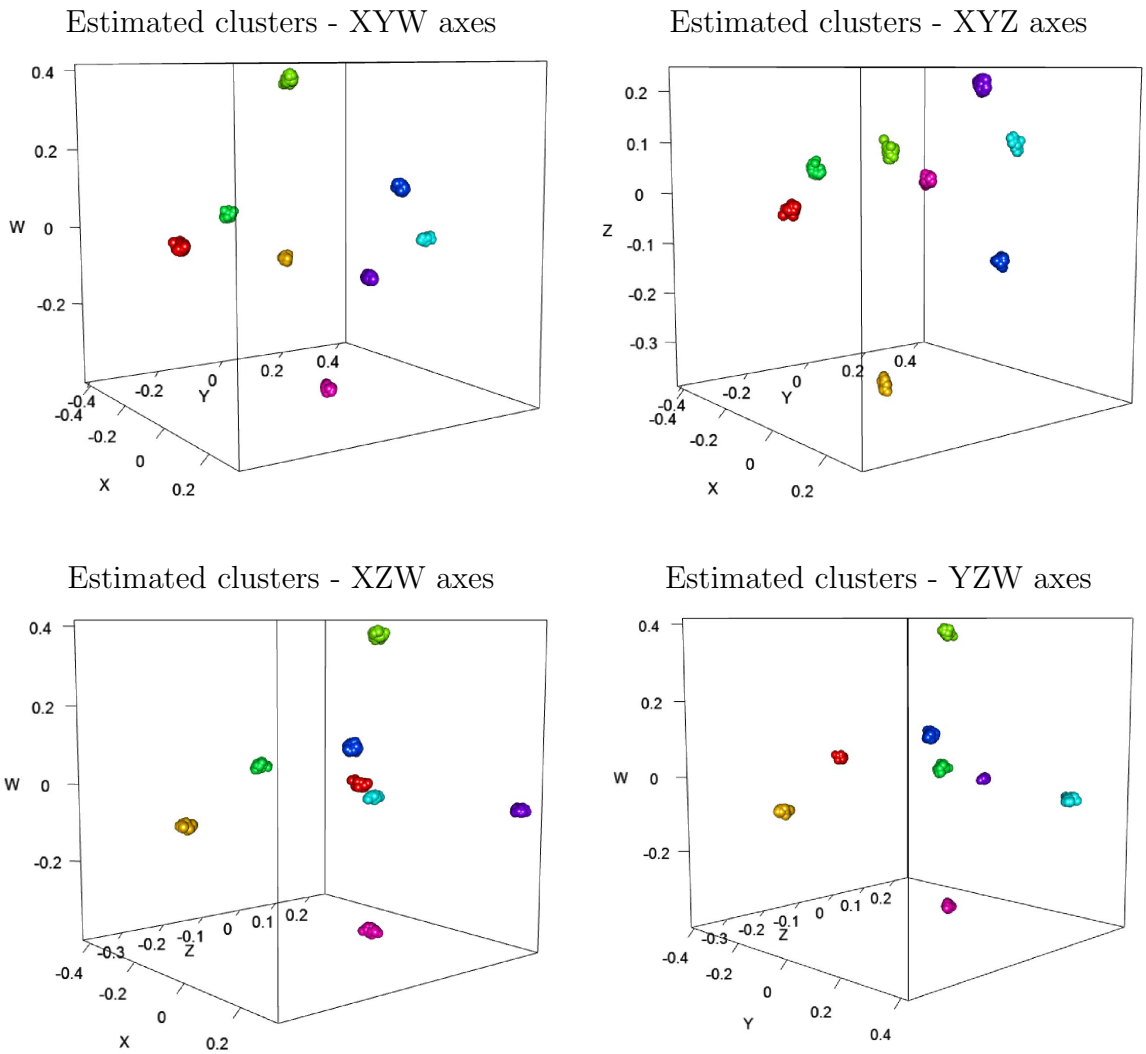



Figure S13: Snapshots of the four dimensional $l_1$-MDS models showing the estimated clusters constructed by removing one axis at a time.

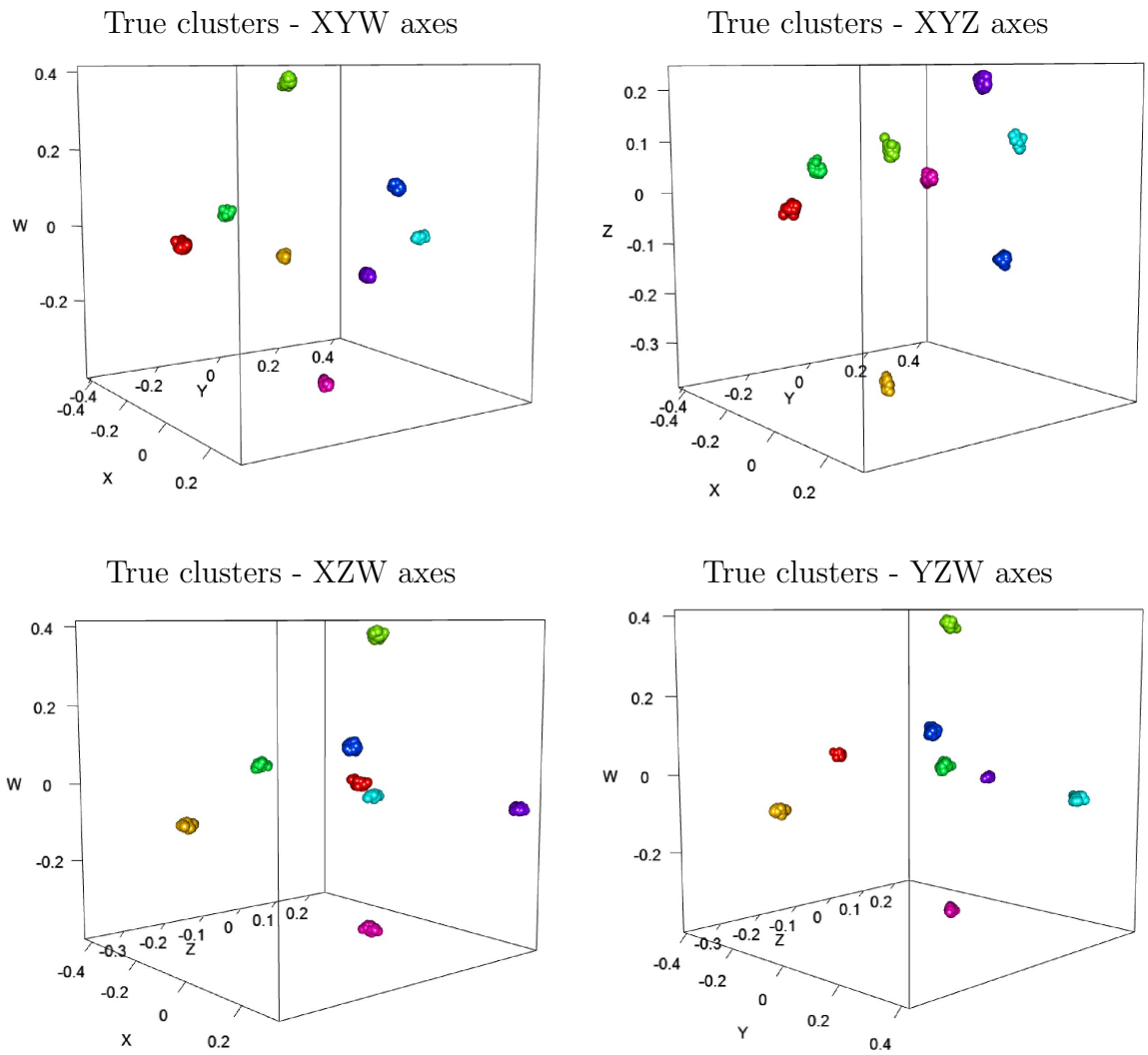



Figure S14: Snapshots of the four dimensional $l_1$-MDS models showing the true clusters constructed by removing one axis at a time.

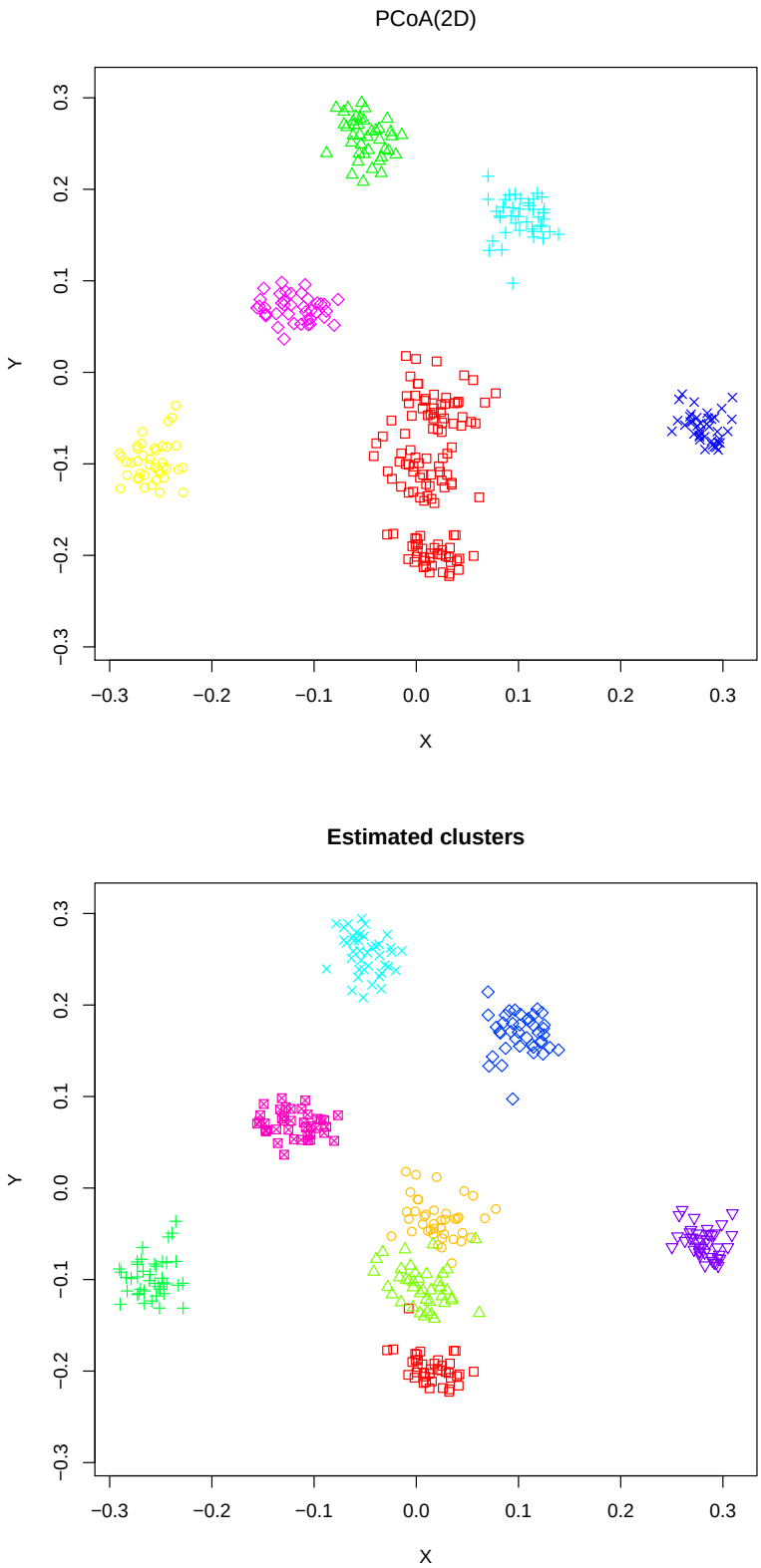
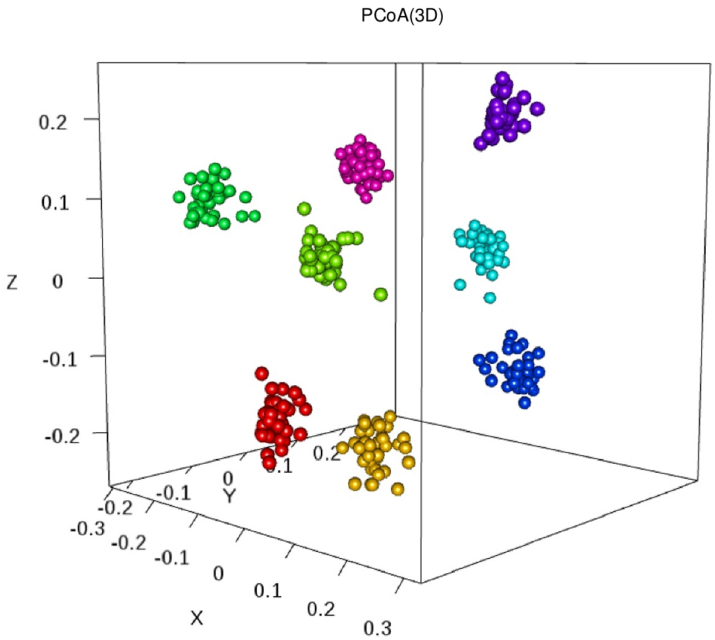



Figure S15: Two dimensional PCoA models displaying the estimated and true clusters for the generated data. The estimated number of clusters is 4 and the true number is equal to 10.

Estimated clusters




True clusters

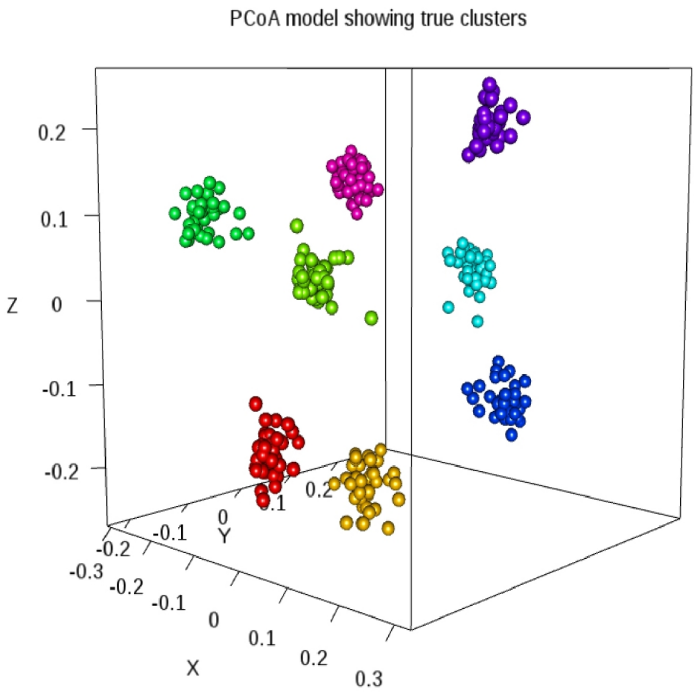



Figure S16: Snapshots of three dimensional PCoA models displaying the estimated and true clusters. Interactive models can be opened by clicking on the links provided above the graph.

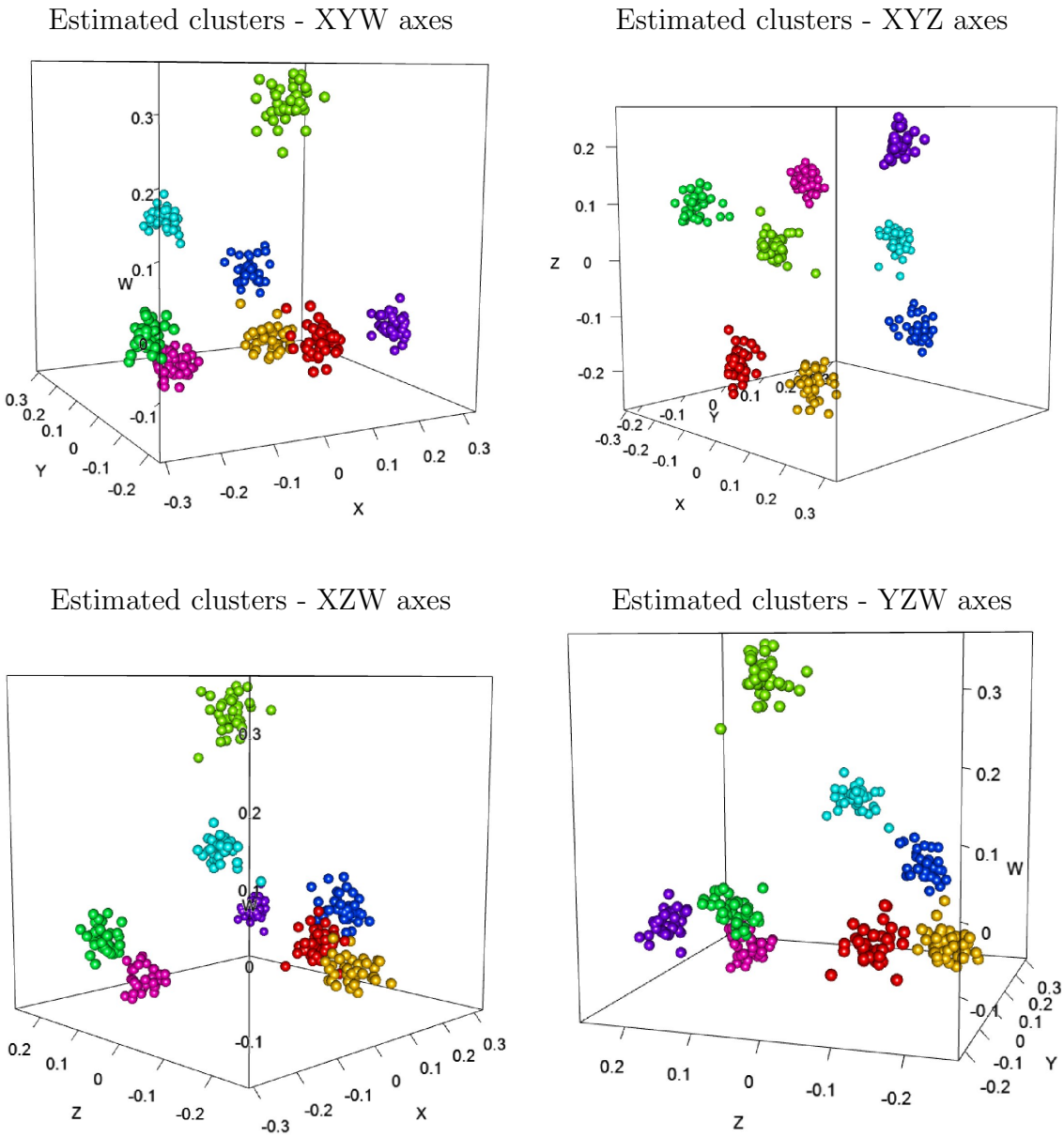



Figure S17: Snapshots of the four dimensional PCoA models showing the estimated clusters constructed by removing one axis at a time.

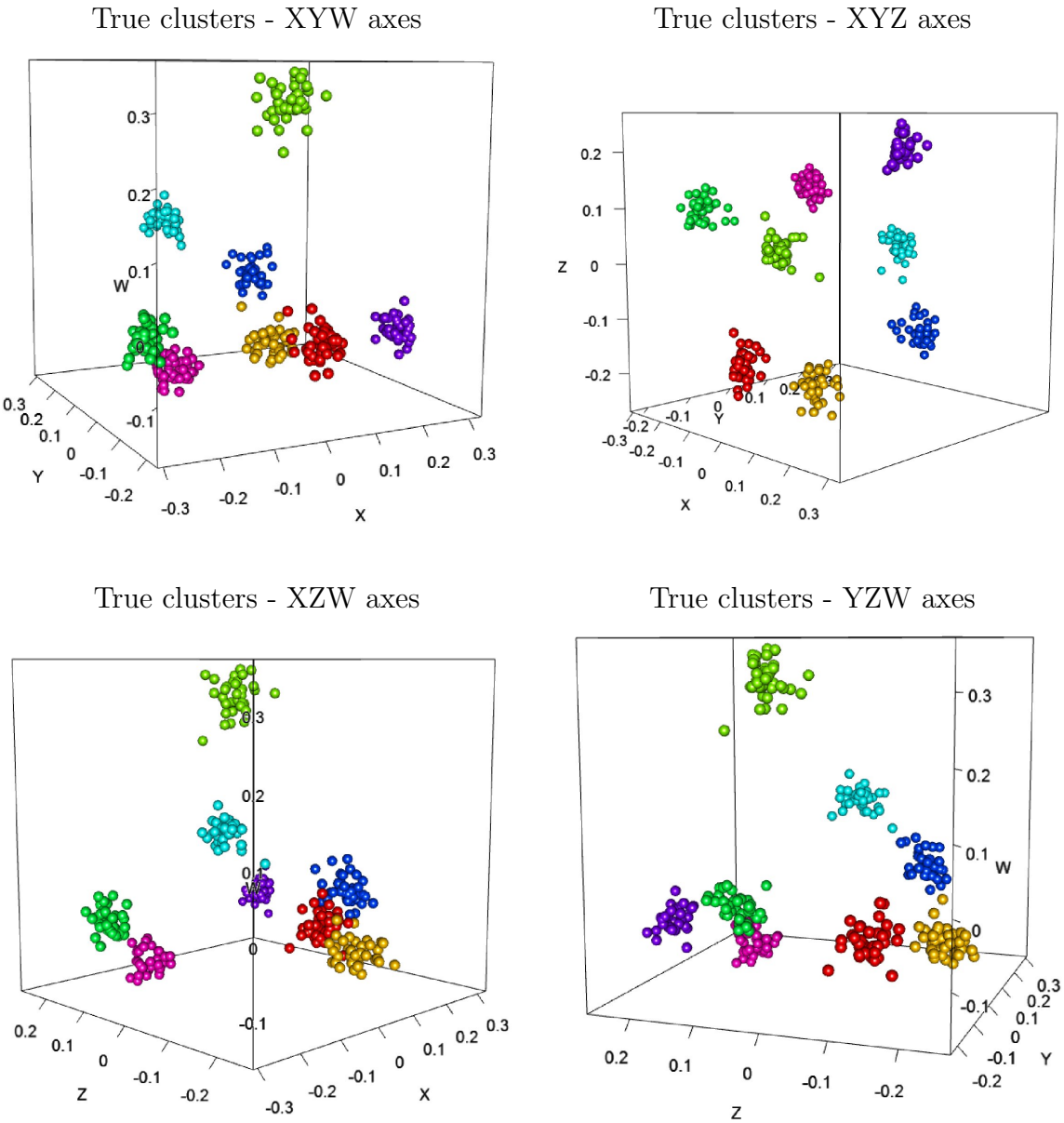



Figure S18: Snapshots of the four dimensional PCoA models showing the true clusters constructed by removing one axis at a time.

# 3 Costello et al (2009) data

## 3.1 Silhouette plot

Preliminary study of the data revealed a clear separation of the data into two clusters, one of which corresponds to all samples from gut. Therefore, to study further clustering in the data, we considered estimation of the optimal number of clusters assuming that the quantity is at least 3. From the silhouette plot presented in Figure S19, the optimal number of clusters for the MDS models of dimension 2, 3 and 4 are 5, 7 and 11 respectively. For the PCoA models, the numbers are 3, 3 and 6 respectively.

## 3.2 Clustering taxa for gut samples

To demonstrate the utility of the GrammR package, we studied the gut samples from the [Costello *et al*, 2009] data. For the 45 samples collected from the gut, metagenomic counts were reported with family as the taxonomic ranking. A total of 312 taxa at the bacterial family level are recorded. Of these 312 taxa, 257 taxa did not have any recorded sequences classified into them, resulting in zero count for all the samples for these taxa. Hence, these taxa are not of significance and are not considered for the analysis. We then analyzed the remaining 55 taxa by constructing $l_1$-MDS of dimension 2 as an illustration. The optimal number r of clusters is estimated to be 4. The true cluster membership of taxa has not been established, therefore misclassification error(MCE) cannot be calculated. Graphical representation showing the estimated clusters are presented in Figures S20. From the figure, one can see that the $l_1$-MDS model has four clusters with 12, 16, 11 and 16 taxa respectively within each cluster.

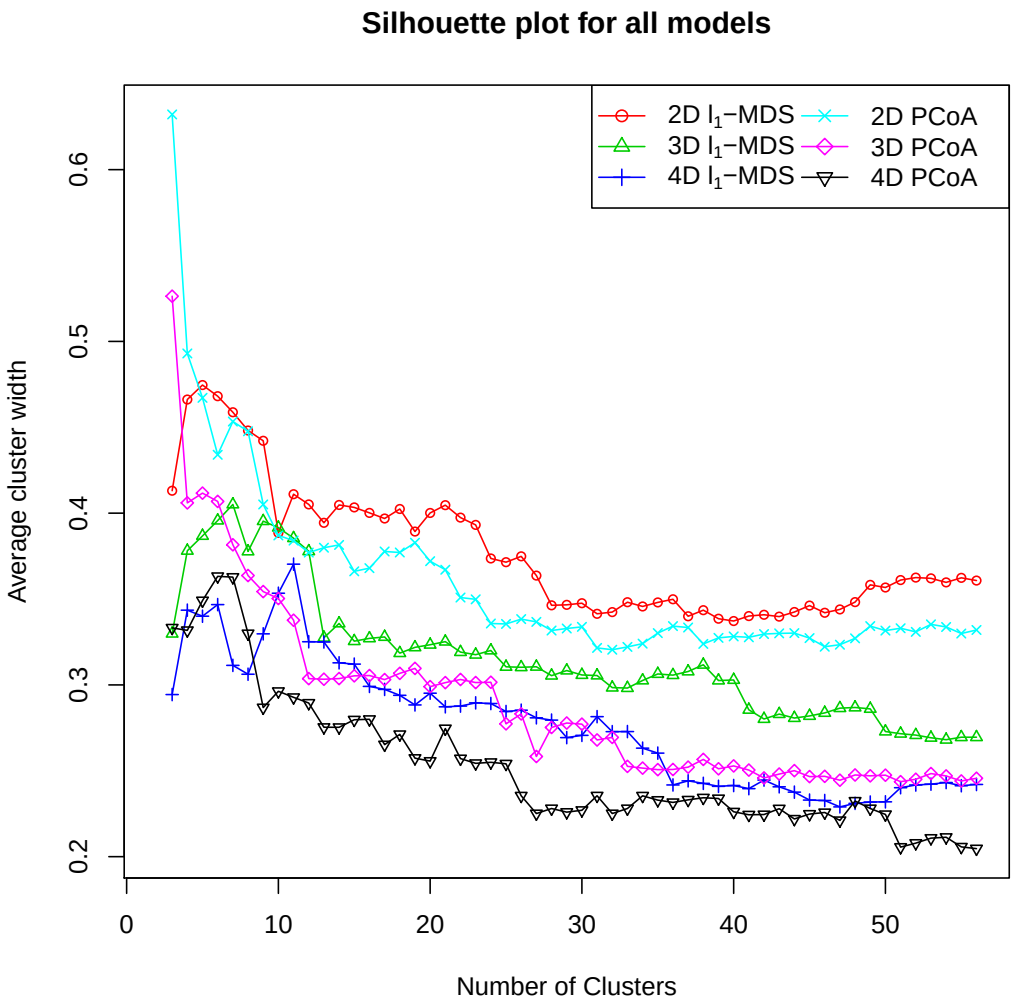



Figure S19: Average silhouette plot of all the six models for Costello et al data.

# 4 Software and example

GrammR is implemented as an R-package available at `http://www.stat.osu.edu/~statgen/SOFTWARE/GrammR/`. It may also be downloaded from `http://cran.rproject.org/web/packages/GrammR/`. For running the GUI version, example data are available at `http://www.stat.osu.edu/~statgen/SOFTWARE/GrammR/`, so are the files needed for viewing the interactive plots.

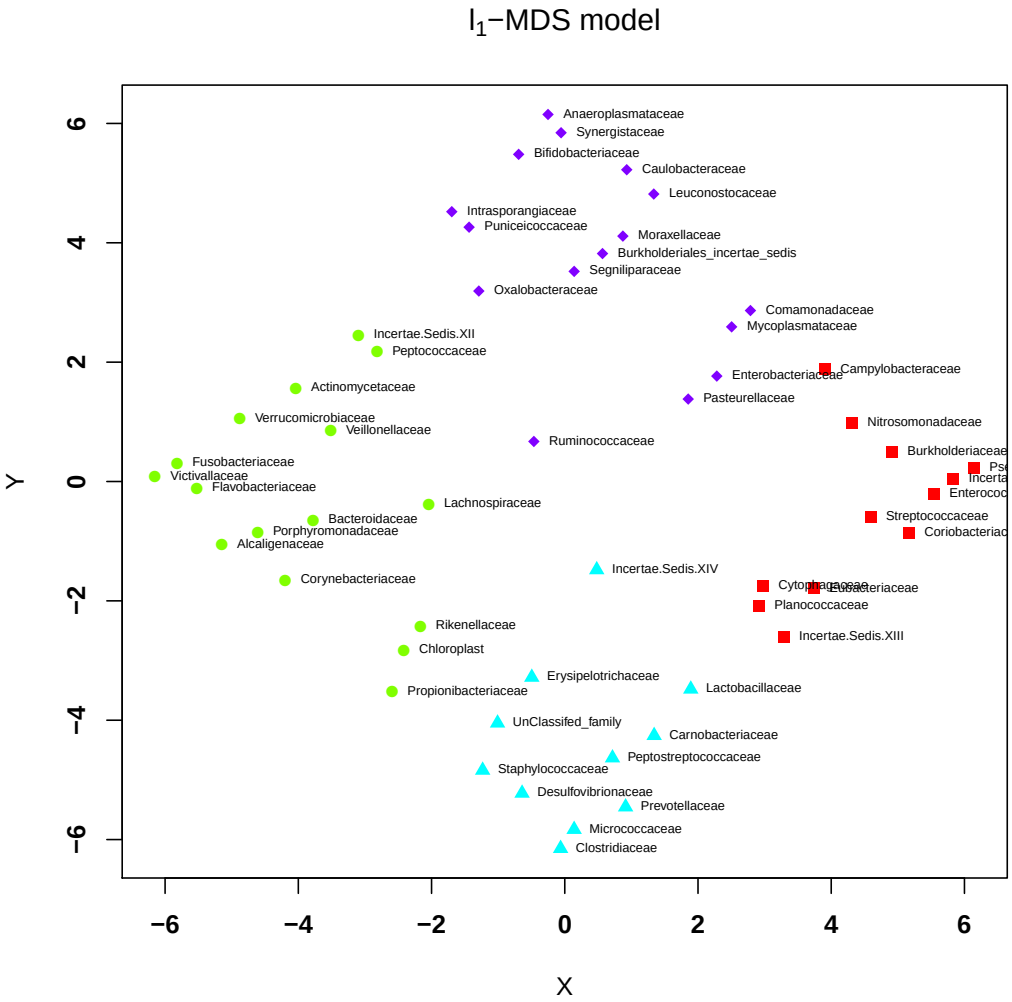



Figure S20: Two dimensional $l_1$-MDS models displaying the estimated clusters for taxa using the gut samples. Name of the taxa corresponding each sample point are provided as labels.

# References

[Costello *et al*, 2009] Costello, Elizabeth K. *et al.* (2009) Bacterial Community Variation in Human Body Habitats Across Space and Time, *Science*, **326**, 1694–1697