Clustering four designs

Contents

1 How the time series plot of the designs look like for different sample sizes?

```
set.seed(9999)
nx_val = 2 # number of x-axis levels
nfacet_val = 3 # number of facet levels
w1_val = 3 # increment in mean
w2_val = 0 # increment in sd
mean_val = 0 # mean of normal distribution of starting combination
sd_val = 1 # sd of normal distribution of starting combination
quantile_prob_val = seq(0.1, 0.9, 0.1)
```

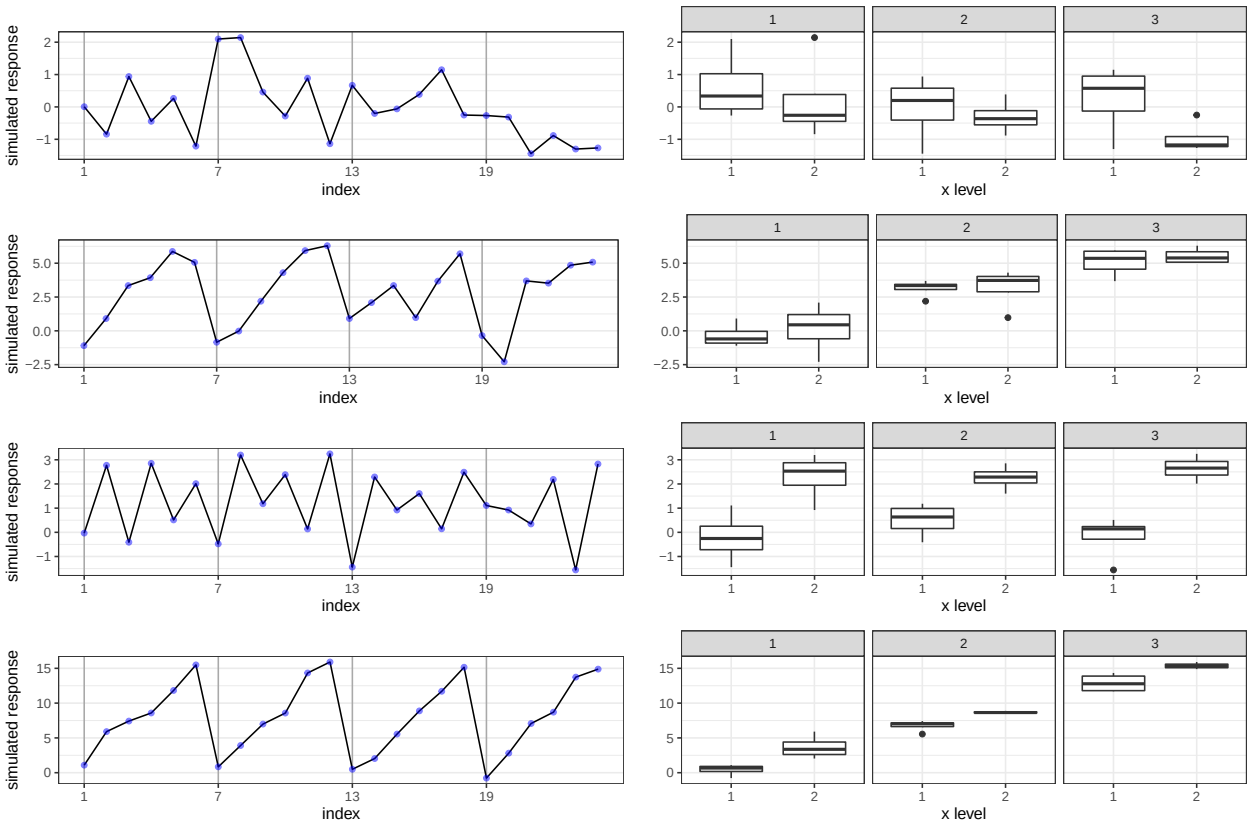
1.1 nobs per combination is 4

```
ntimes_val = 4
```

Table 1: Simulated data across combination of categories (left) and manipulated time series format (right)

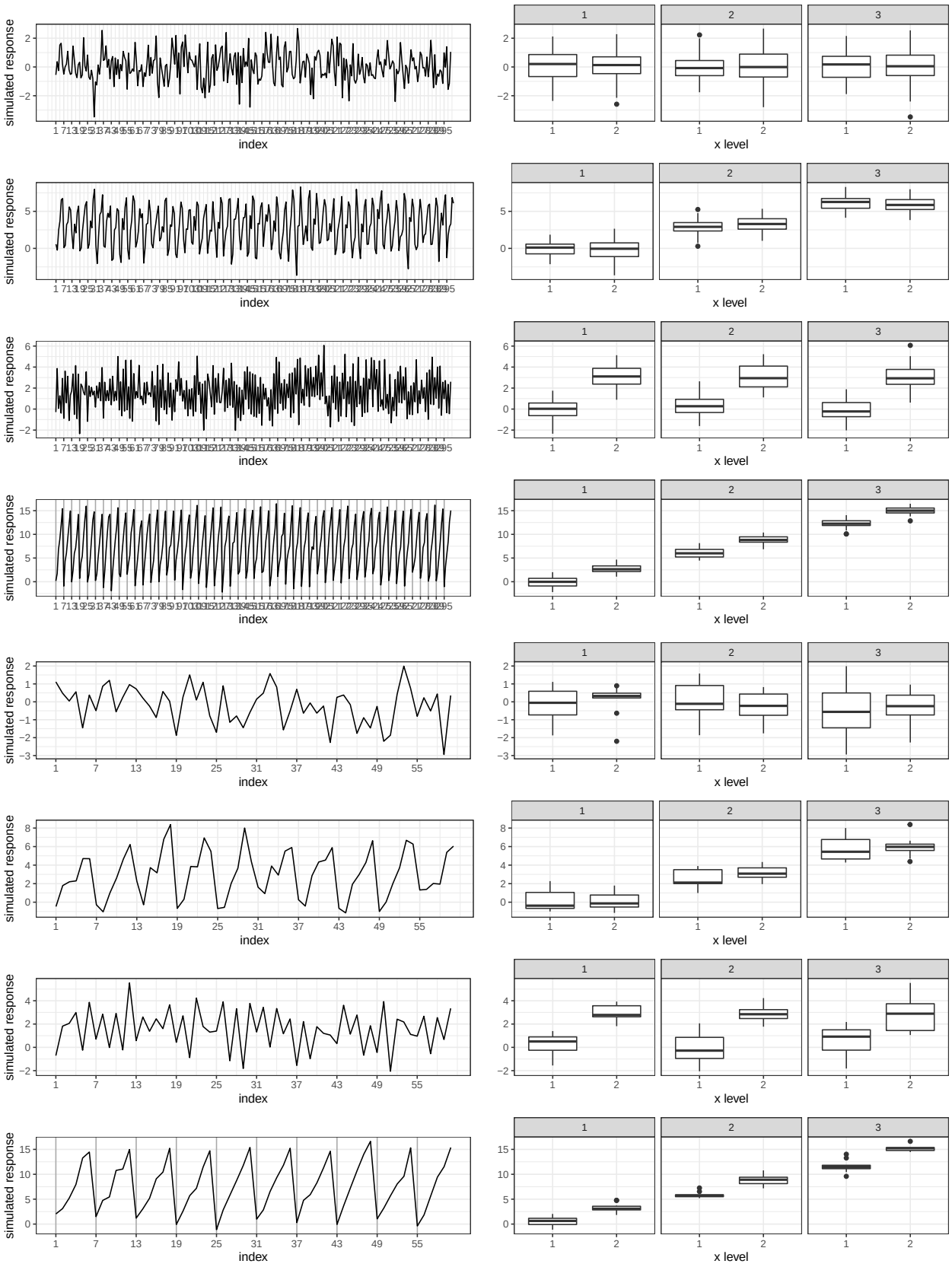
id_facet	id_x	sim_data
1	1	1.0840991
1	1	0.8431089
1	1	0.4943890
1	1	-0.7730161
1	2	5.9038161
1	2	3.9088839
1	2	2.0369233
1	2	2.8117113
2	1	7.4203744
2	1	6.9864382
2	1	5.5500595
2	1	7.0716663
2	2	8.5834288
2	2	8.5708607
2	2	8.8833411
2	2	8.7014611
3	1	11.8207616
3	1	14.3329026
3	1	11.7010540
3	1	13.7342014
3	2	15.4918386
3	2	15.9084482
3	2	15.1561588
3	2	14.8834948

id_facet	id_x	sim_data	index_old	index_new	time
1	1	1.0840991	1	1	1
1	2	5.9038161	5	17	2
2	1	7.4203744	9	10	3
2	2	8.5834288	13	3	4
3	1	11.8207616	17	19	5
3	2	15.4918386	21	12	6
1	1	0.8431089	2	5	7
1	2	3.9088839	6	21	8
2	1	6.9864382	10	14	9
2	2	8.5708607	14	7	10
3	1	14.3329026	18	23	11
3	2	15.9084482	22	16	12
1	1	0.4943890	3	9	13
1	2	2.0369233	7	2	14
2	1	5.5500595	11	18	15
2	2	8.8833411	15	11	16
3	1	11.7010540	19	4	17
3	2	15.1561588	23	20	18
1	1	-0.7730161	4	13	19
1	2	2.8117113	8	6	20
2	1	7.0716663	12	22	21
2	2	8.7014611	16	15	22
3	1	13.7342014	20	8	23
3	2	14.8834948	24	24	24



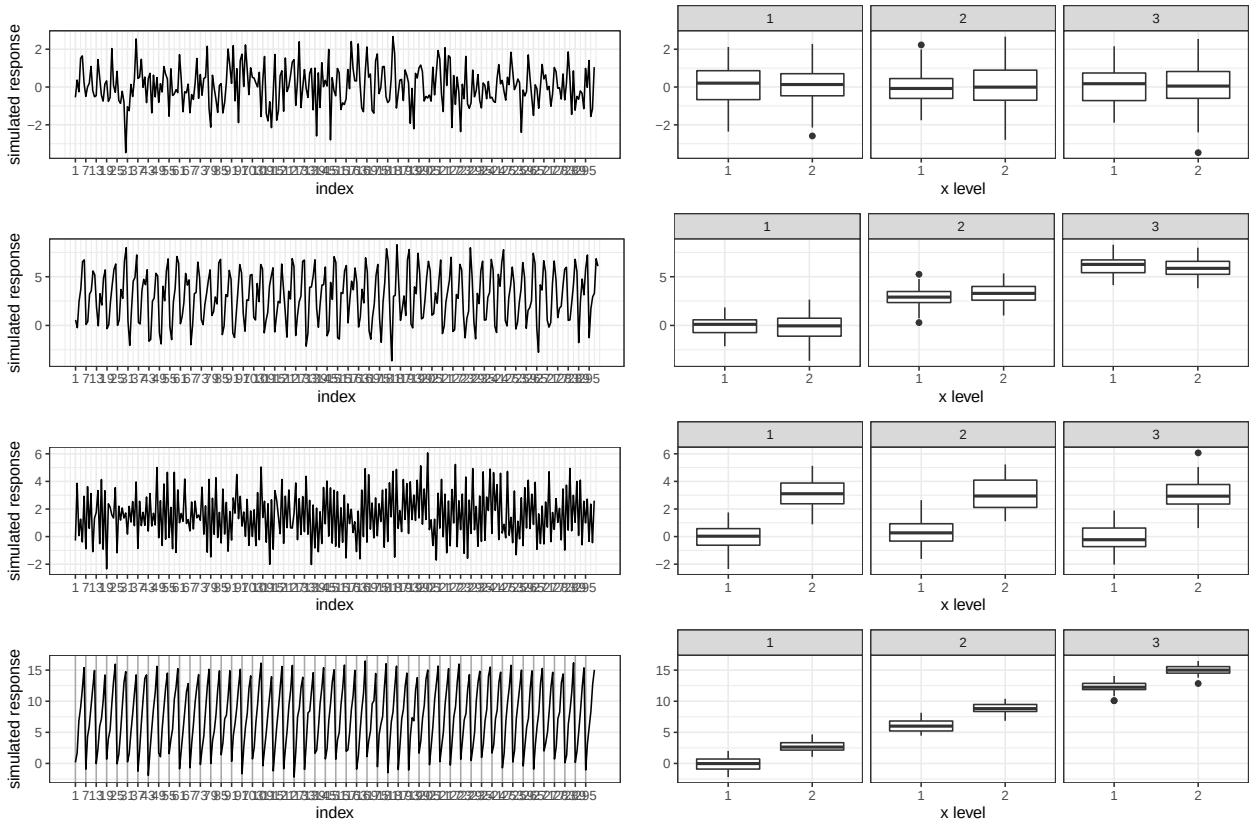
1.2 nobs per combination is 10

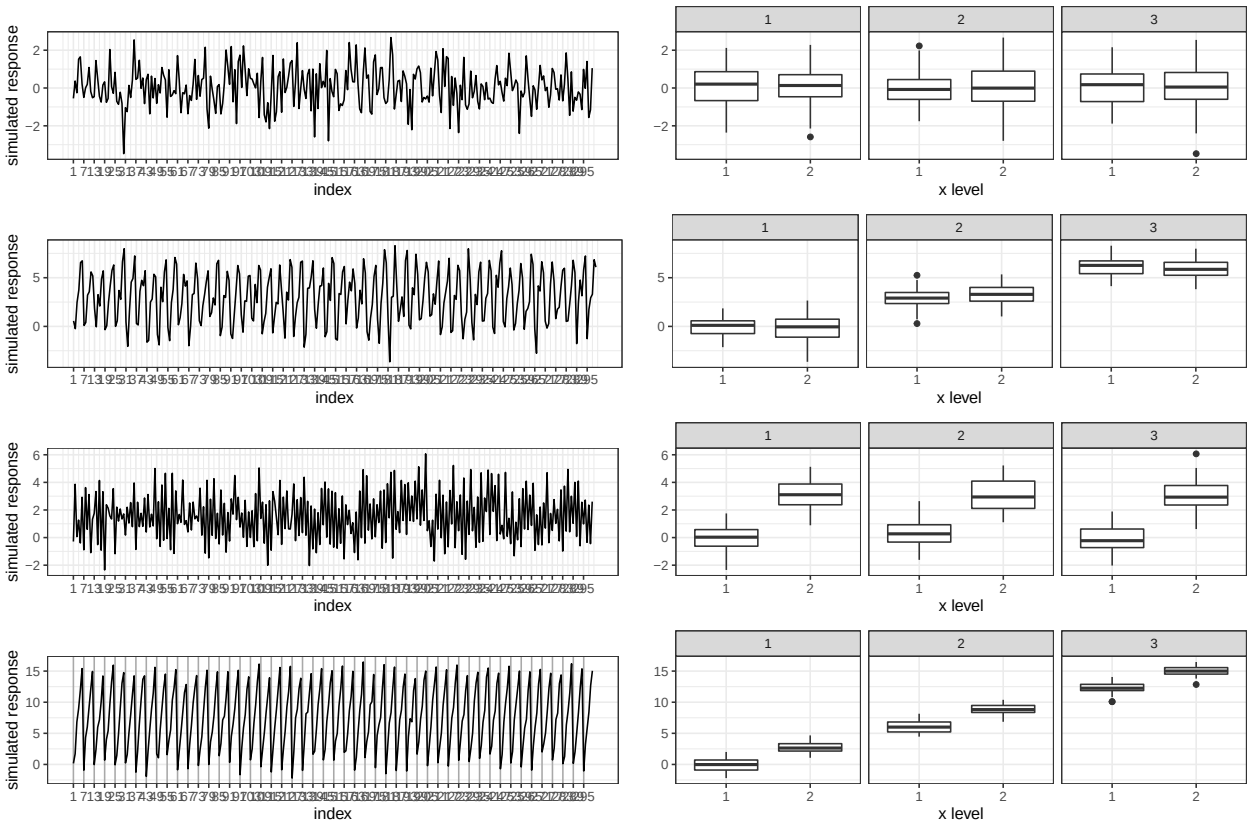
```
ntimes_val = 10 # nobs per combination
```


simulated response
index
x level
1
2
3

1.3 nobs per combination is 50

```
ntimes_val = 50 # nobs per combination
```





2 Clustering on simulated datasets

```
sample_seed <-  seq(1, 100, 1)
```

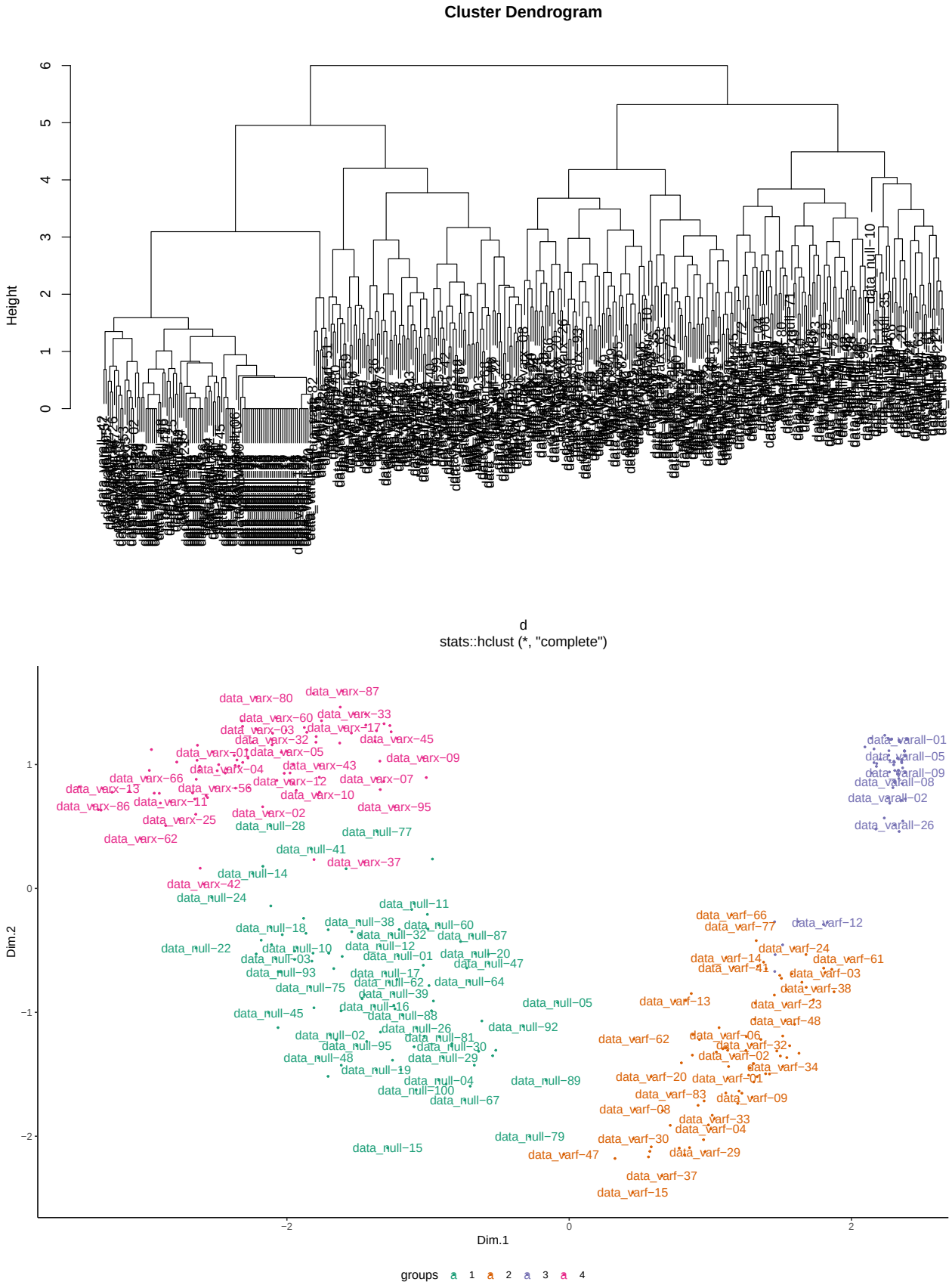
DGP: Generate 10 time series from each designs Time series are simulated from each of these designs with 50 observations in each group. So we have 50 observations each for the six combination of categories $(1,1),(1,2),(2,1),(2,2),(3,1),(3,2)$. Time series are simulated for ten different seeds (from `sample_seed`) for each design. `data_varall-1` represents a data set from a design $D_{var_{all}}$ (distributions change across both facet and x) with a seed 10. `data_varall-2` represents a data set from a design $D_{var_{all}}$ with a seed 20 and so on. `data_null`, `data_varx` and `data_varf`corresponds to designs D_{null}, D_{var_x} and D_{var_f} designs respectively.

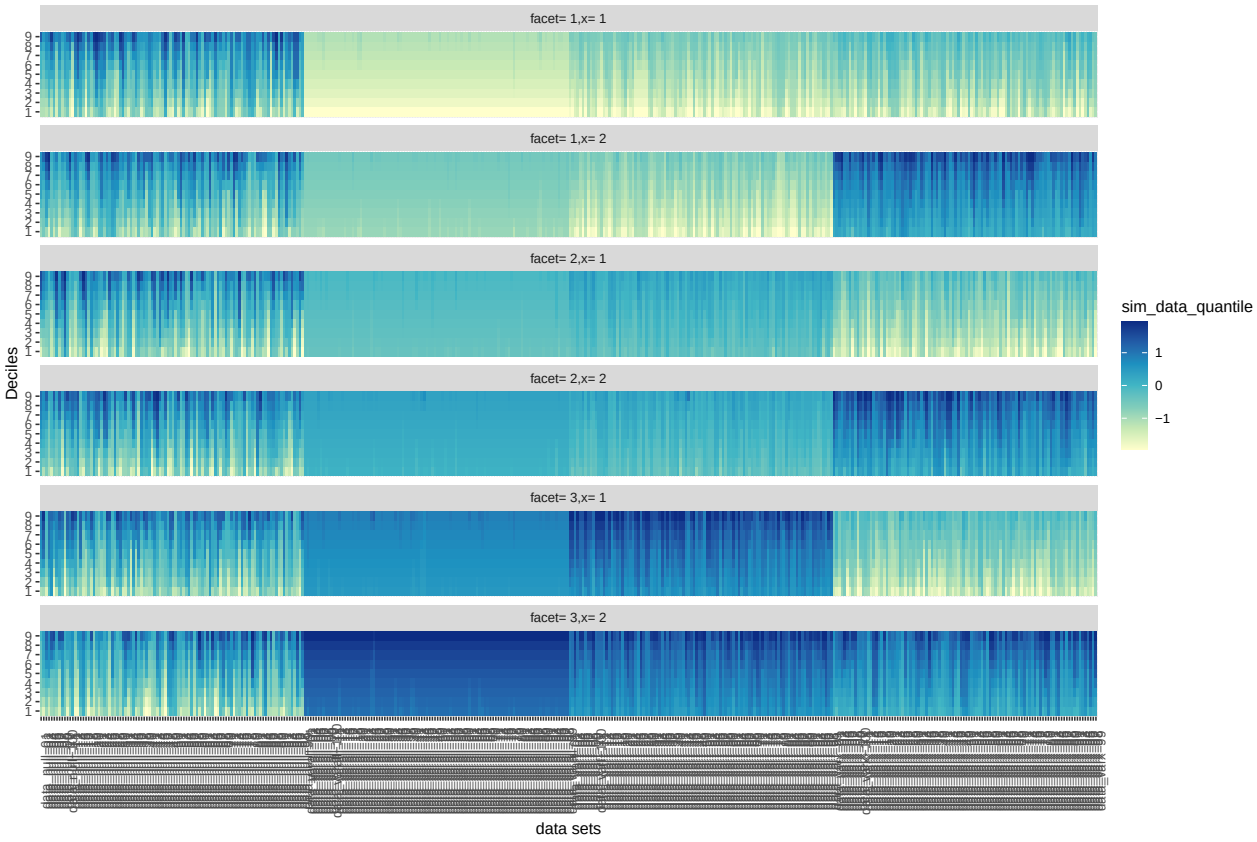
Compute quantiles of conditional distributions Conditional quantiles are obtained for each combination of categories.

JS Pairwise distances between datasets Distance between the data sets is computed as the sum of JS distances across different categories.

Hierarchical clustering with 4 clusters Hierarchical clustering is performed using k = 4 and dendogram observed

Clusters obtained visualized using MDS Each cluster represents data sets from a separate design


Cluster Dendrogram
Height
0
1
2
3
4
5
6
d
stats::hclust (*, "complete")
Dim.2
Dim.1
−2
−1
0
1
2
groups
1
2
3
4



2.1 Cluster validation

```
## $n
## [1] 400
##
## $cluster.number
## [1] 4
##
## $cluster.size
## [1] 100  93 107 100
##
## $min.cluster.size
## [1] 93
##
## $noisen
## [1] 0
##
## $diameter
## [1] 4.490923 4.204883 3.093958 4.179376
##
## $average.distance
## [1] 2.7301025 2.1335705 0.7789936 2.4061209
##
## $median.distance
## [1] 2.7240339 2.0907343 0.5809741 2.3814645
##
## $separation
```

```
## [1] 2.115852 1.176403 1.176403 2.115852
##
## $average.toother
## [1] 4.195385 4.004869 4.290831 4.468051
##
## $separation.matrix
##          [,1]     [,2]     [,3]     [,4]
## [1,] 0.000000 2.559864 3.021873 2.115852
## [2,] 2.559864 0.000000 1.176403 2.997538
## [3,] 3.021873 1.176403 0.000000 3.181885
## [4,] 2.115852 2.997538 3.181885 0.000000
##
## $ave.between.matrix
##          [,1]     [,2]     [,3]     [,4]
## [1,] 0.000000 4.166543 4.600338 3.788907
## [2,] 4.166543 0.000000 3.252935 4.647764
## [3,] 4.600338 3.252935 0.000000 4.946566
## [4,] 3.788907 4.647764 4.946566 0.000000
##
## $average.between
## [1] 4.243198
##
## $average.within
## [1] 1.988492
##
## $n.between
## [1] 59951
##
## $n.within
## [1] 19849
##
## $max.diameter
## [1] 4.490923
##
## $min.separation
## [1] 1.176403
##
## $within.cluster.ss
## [1] 952.1591
##
## $clus.avg.silwidths
##         1         2         3         4
## 0.2484767 0.3222798 0.7438105 0.3649062
##
## $avg.silwidth
## [1] 0.4272451
##
## $g2
## NULL
##
## $g3
## NULL
##
## $pearsongamma
```

```
## [1] 0.7669777
##
## $dunn
## [1] 0.2619512
##
## $dunn2
## [1] 1.191506
##
## $entropy
## [1] 1.385068
##
## $wb.ratio
## [1] 0.4686305
##
## $ch
## [1] 287.4193
##
## $cwidegap
## [1] 2.191903 1.606886 1.589187 2.001813
##
## $widestgap
## [1] 2.191903
##
## $sindex
## [1] 1.5088
##
## $corrected.rand
## [1] 0.9546631
##
## $vi
## [1] 0.1280455
```

The total average (mean of all individual silhouette widths) is 0.427 and corrected Rand index is 0.955