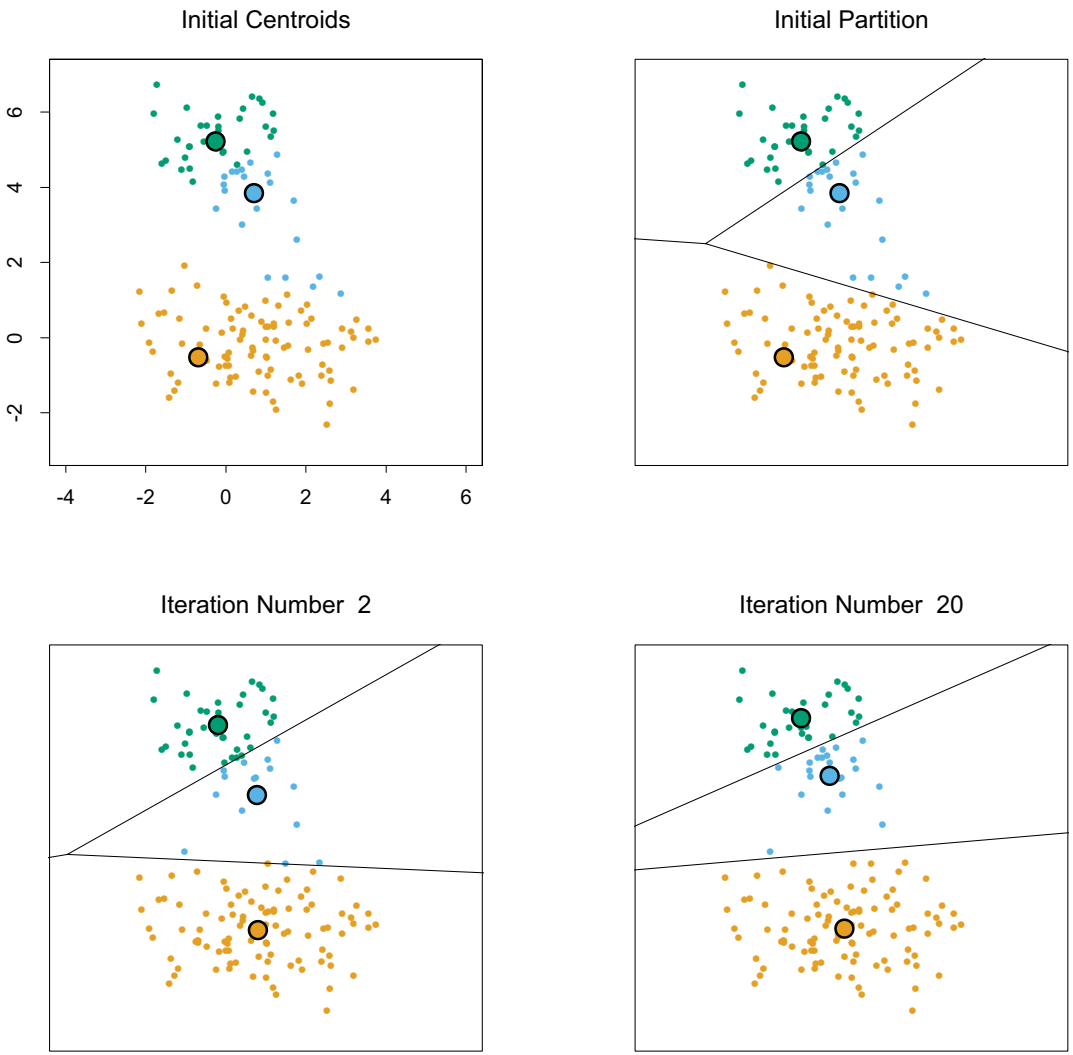
Clustering

Data Mining
Prof. Dawn Woodard
School of ORIE
Cornell University

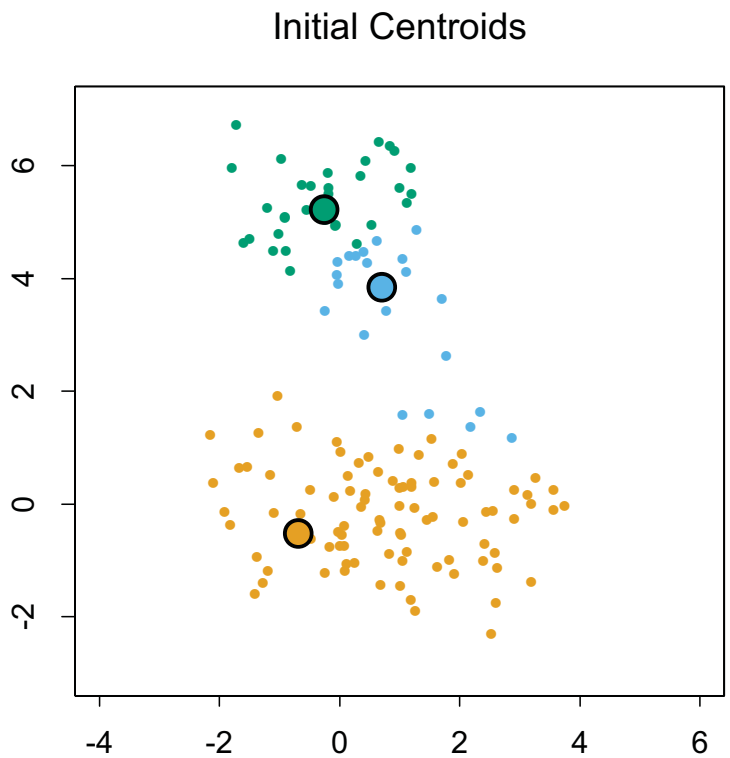
Outline

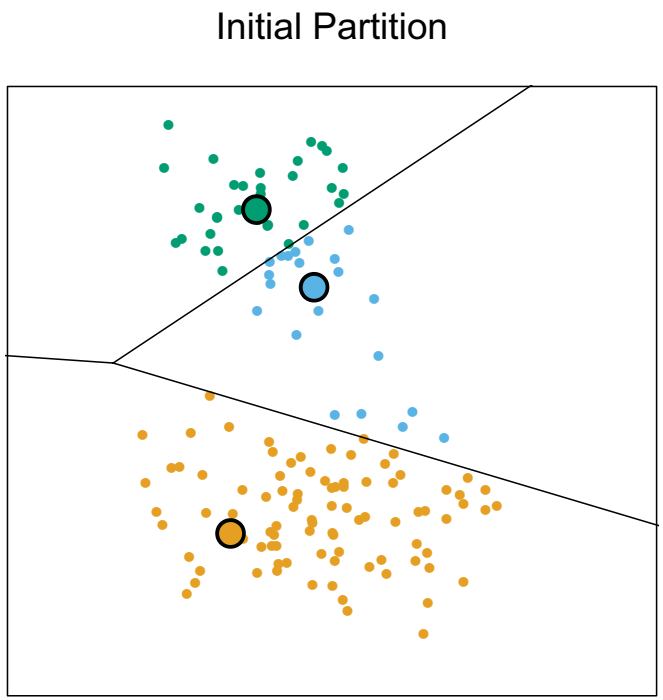
K-Means Clustering

Figure 14.6 in the Hastie et al. text:



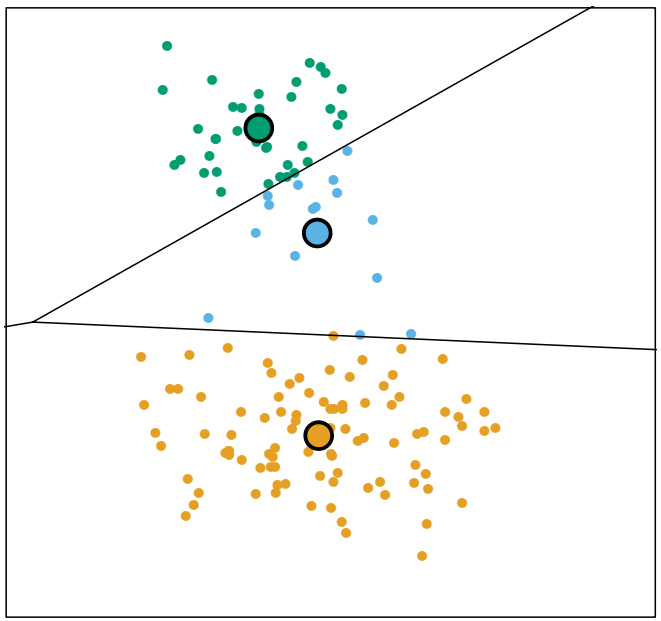
K-Means Clustering



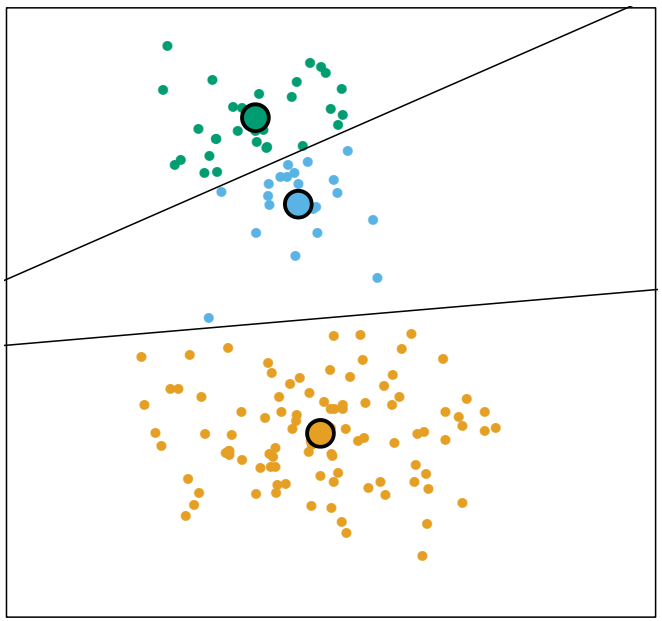


K-Means Clustering

Iteration Number 2



Iteration Number 20



Public Utilities Data

- Recall the data on public utilities (Shmueli, Patel, and Bruce, 2007)
- Wish to group based on financial factors
- Used for e.g. a study on the impact of deregulation
- Could pick one "typical" utility in each group and study in detail the potential effect of deregulation on that utility
- Scale up to estimate impact for all utilities
- This is less costly than studying in detail the effect of deregulation on every single utility

Utilities Data

Data on public utilities (Shmueli, Patel, and Bruce, 2007):

	Company	Fixed.charge	RoR	Cost	Load.factor	Demand.growth	Sales	Nuclear	Fuel.Cost
1	Arizona	1.06	9.2	151	54.4	1.6	9077	0.0	0.628
2	Boston	0.89	10.3	202	57.9	2.2	5088	25.3	1.555
3	Central	1.43	15.4	113	53.0	3.4	9212	0.0	1.058
4	Commonwealth	1.02	11.2	168	56.0	0.3	6423	34.3	0.700
5	NY	1.49	8.8	192	51.2	1.0	3300	15.6	2.044
6	Florida	1.32	13.5	111	60.0	-2.2	11127	22.5	1.241
7	Hawaiian	1.22	12.2	175	67.6	2.2	7642	0.0	1.652
8	Idaho	1.10	9.2	245	57.0	3.3	13082	0.0	0.309
9	Kentucky	1.34	13.0	168	60.4	7.2	8406	0.0	0.862
10	Madison	1.12	12.4	197	53.0	2.7	6455	39.2	0.623
11	Nevada	0.75	7.5	173	51.5	6.5	17441	0.0	0.768
12	New England	1.13	10.9	178	62.0	3.7	6154	0.0	1.897
13	Northern	1.15	12.7	199	53.7	6.4	7179	50.2	0.527
14	Oklahoma	1.09	12.0	96	49.8	1.4	9673	0.0	0.588
15	Pacific	0.96	7.6	164	62.2	-0.1	6468	0.9	1.400
16	Puget	1.16	9.9	252	56.0	9.2	15991	0.0	0.620
17	San Diego	0.76	6.4	136	61.9	9.0	5714	8.3	1.920
18	Southern	1.05	12.6	150	56.7	2.7	10140	0.0	1.108
19	Texas	1.16	11.7	104	54.0	-2.1	13507	0.0	0.636
20	Wisconsin	1.20	11.8	148	59.9	3.5	7287	41.1	0.702
21	United	1.04	8.6	204	61.0	3.5	6650	0.0	2.116
22	Virginia	1.07	9.3	174	54.3	5.9	10093	26.6	1.306

Utilities Data

Data on public utilities (Shmueli, Patel, and Bruce, 2007):

	Company	Fixed.charge	RoR	Cost	Load.factor	Dema
1	Arizona	1.06	9.2	151	54.4	
2	Boston	0.89	10.3	202	57.9	
3	Central	1.43	15.4	113	53.0	
4	Commonwealth	1.02	11.2	168	56.0	
5	NY	1.49	8.8	192	51.2	
6	Florida	1.32	13.5	111	60.0	
7	Hawaiian	1.22	12.2	175	67.6	
8	Idaho	1.10	9.2	245	57.0	
9	Kentucky	1.34	13.0	168	60.4	
10	Madison	1.12	12.4	197	53.0	
11	Nevada	0.75	7.5	173	51.5	
12	New England	1.13	10.9	178	62.0	
13	Northern	1.15	12.7	199	53.7	
14	Oklahoma	1.09	12.0	96	49.8	
15	Pacific	0.96	7.6	164	62.2	
16	Puget	1.16	9.9	252	56.0	
17	San Diego	0.76	6.4	136	61.9	
18	Southern	1.05	12.6	150	56.7	

Utilities Data

- 8 operational variables:

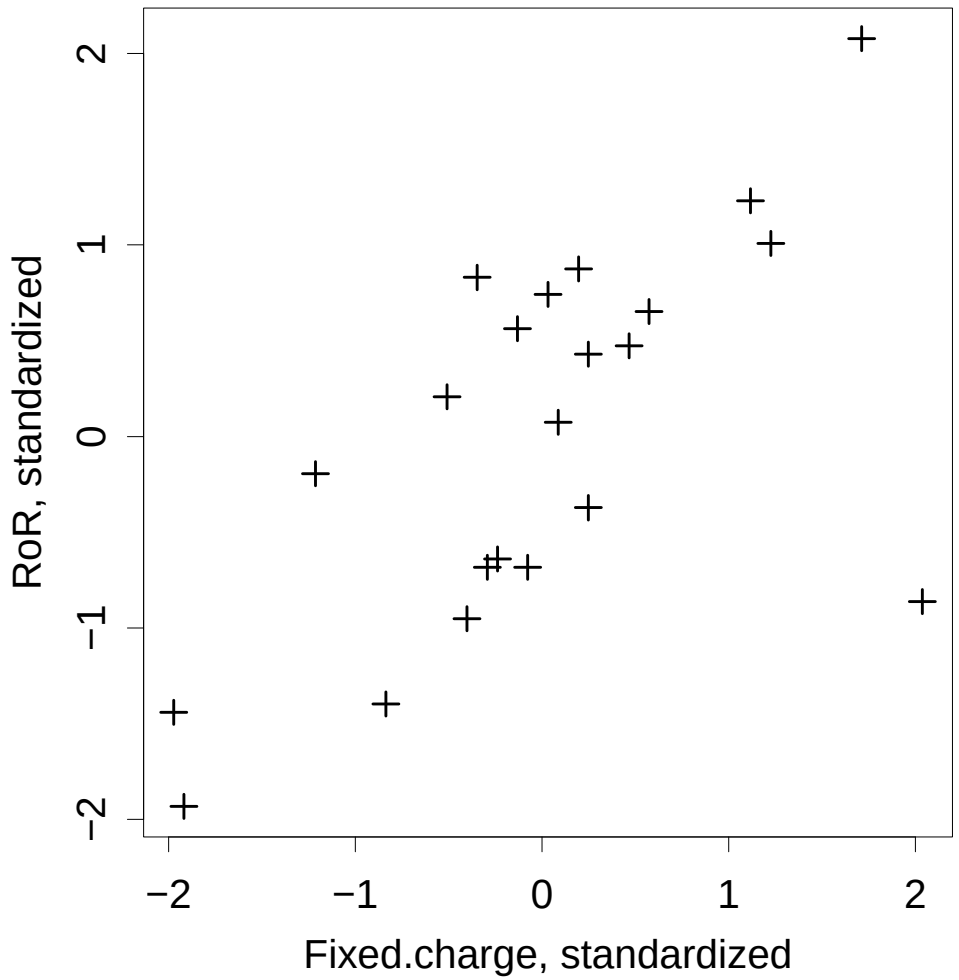
  Fixed.charge: Fixed-charge covering ratio (income/debt)
  RoR: rate of return on capital
  Cost: cost per kilowatt capacity
  Load.factor
  Demand.growth
  Sales: Kilowatthour use per year
  Nuclear: % nuclear
  Fuel.cost: Total fuel costs

Utilities Data

- Let's first cluster based just on the first two variables (Fixed.charge and RoR), so we can visualize the results
- What do you notice about the variance of Fixed.charge relative to that of RoR?
- What effect could that have on the cluster assignments?
- We probably need to standardize the variables

Utilities Data

Plot the resulting variables:

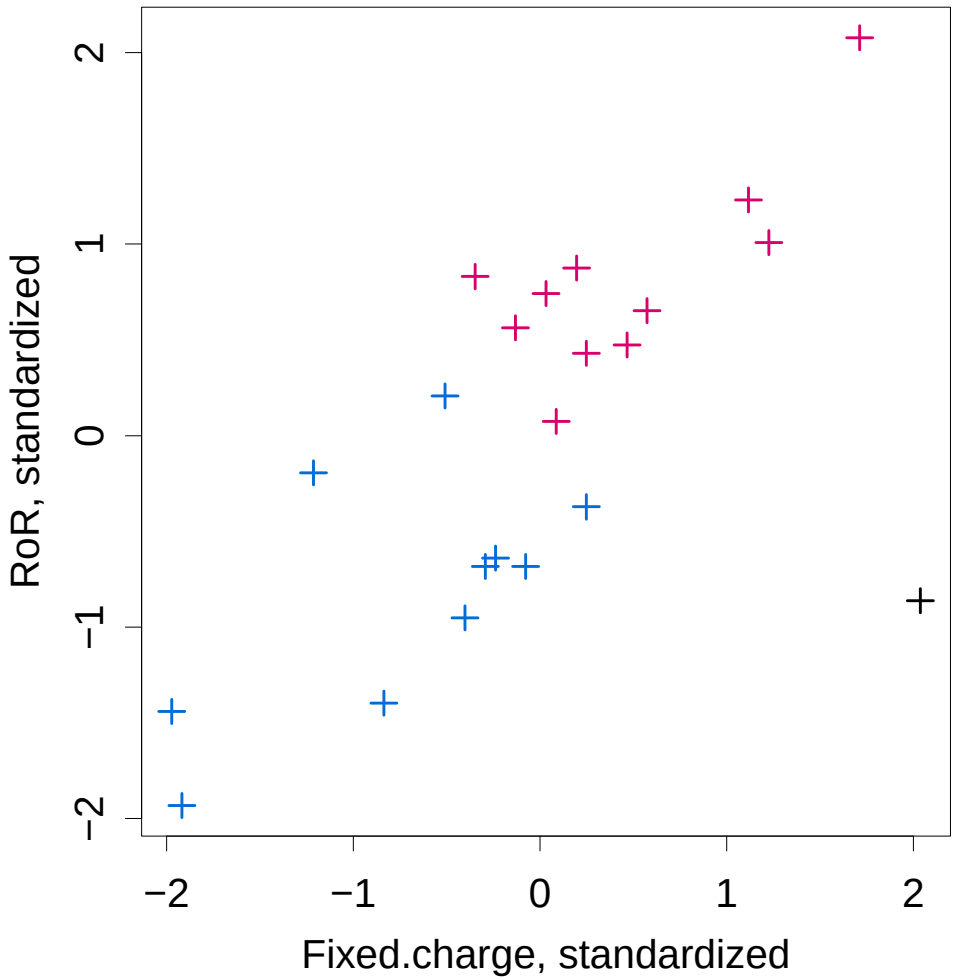


Utilities Data

- What type of relationship do the variables Fixed.charge and RoR appear to have overall?
- Are there any utilities that do not fit this overall trend?
- Are there clear clusters?
- If you had to divide into 3 clusters, what clusters would you use?

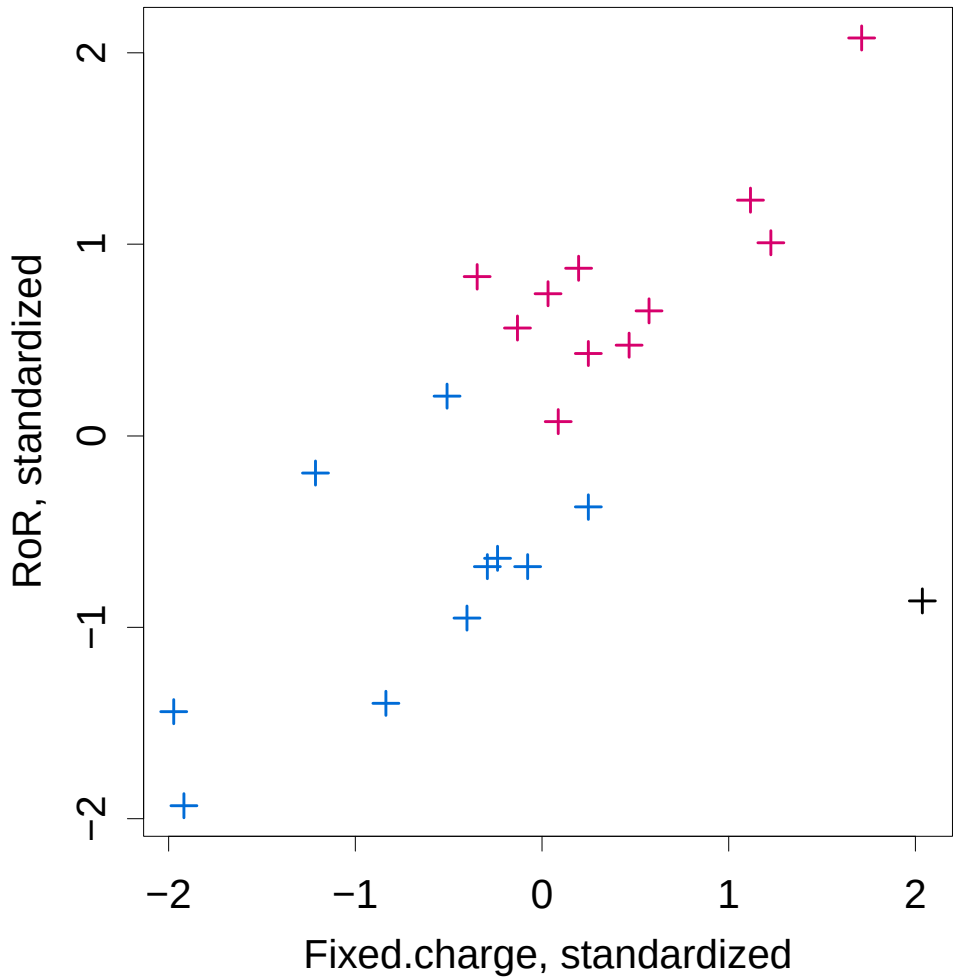
Utilities Data

Applying k-means using one random initialization we obtain the clusters:



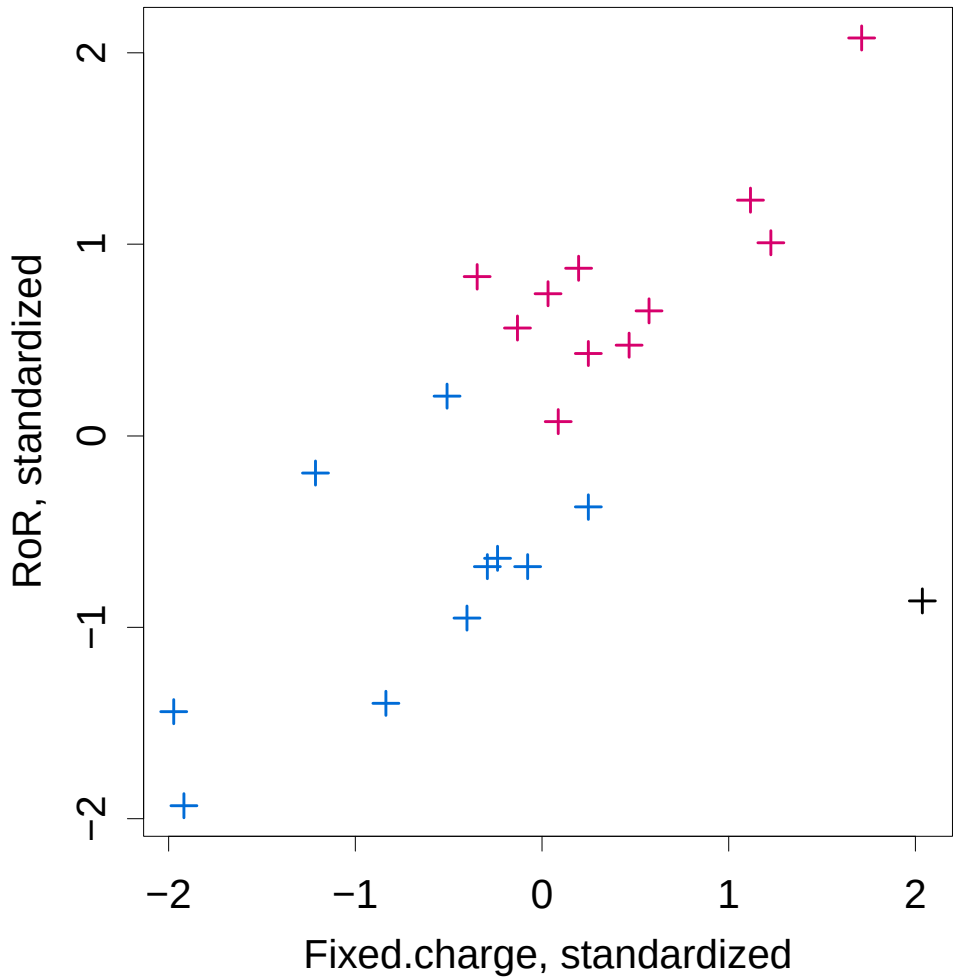
Utilities Data

Here the points are the utilities in the data set



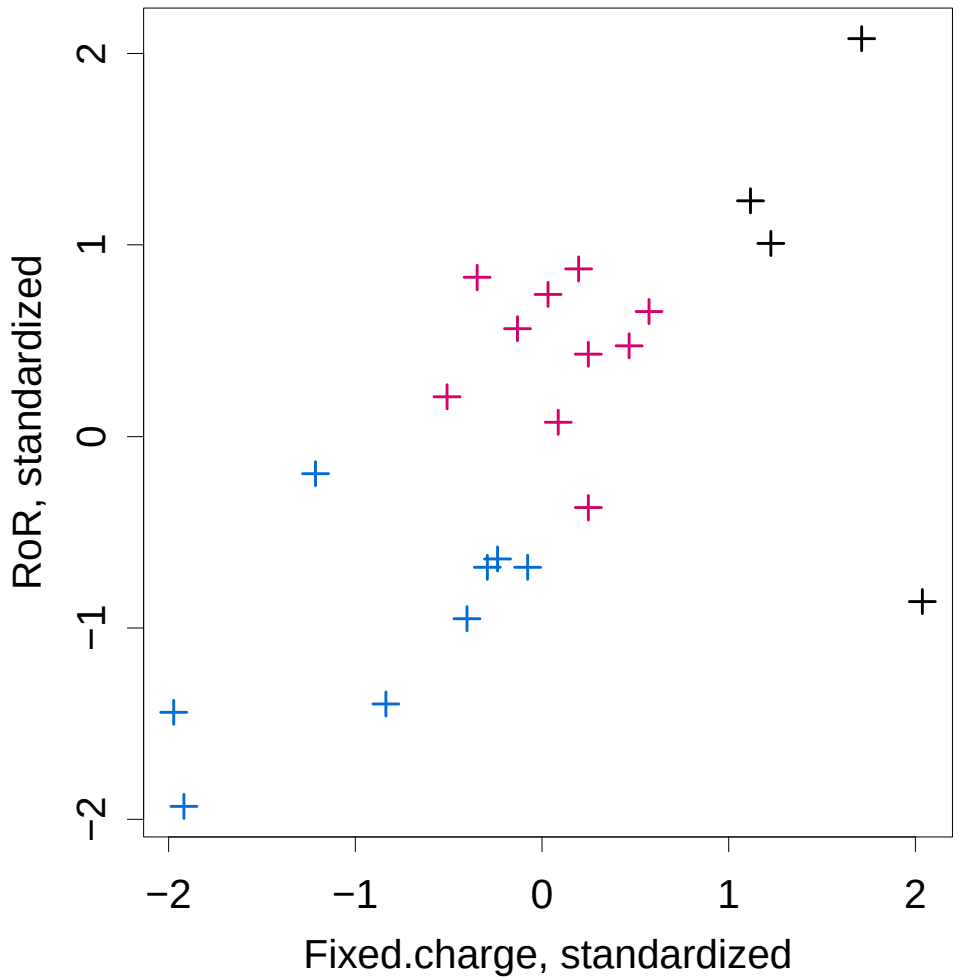
Utilities Data

The colors correspond to the clusters



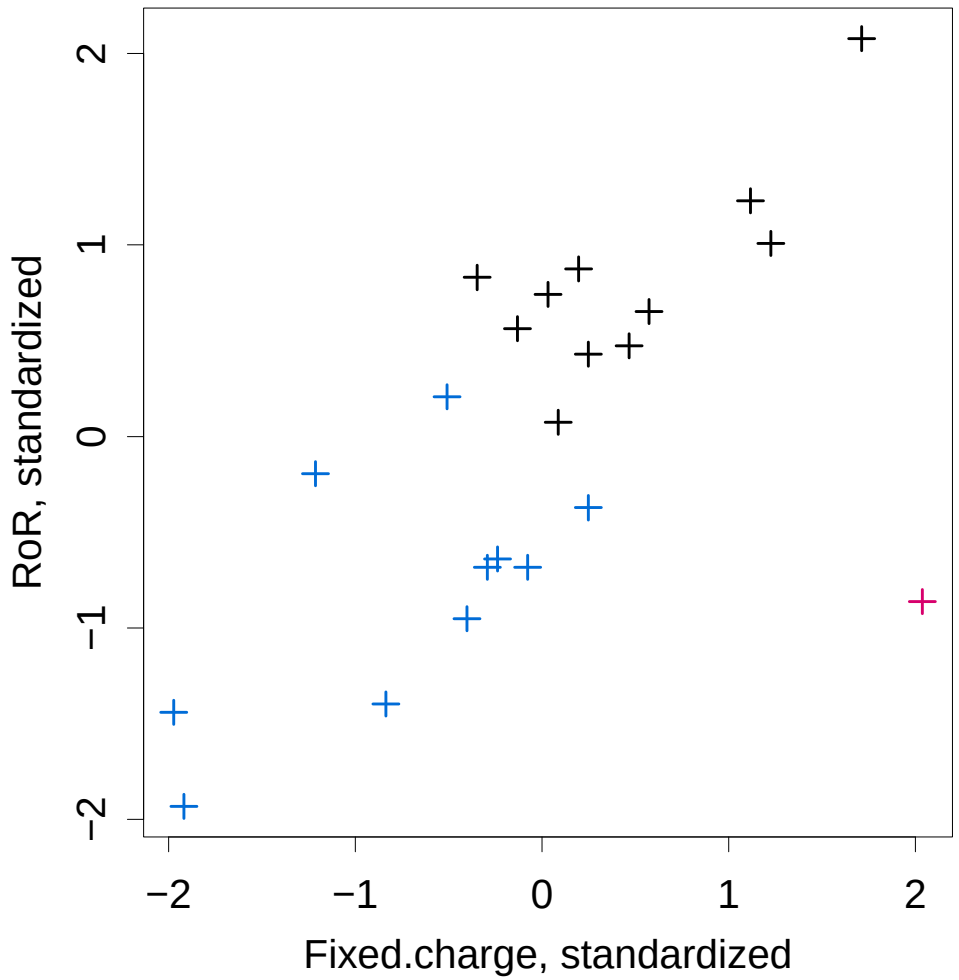
Utilities Data

Using a different random initialization we obtain:



Utilities Data

Using a third random initialization we obtain:



Utilities Data

- How are the first and third cluster assignments related?
- Which cluster assignment do you think is most appropriate?

Utilities Data

- The within-cluster variation $W(C)$ for the first and third cluster assignments is 15.62
- That for the second cluster assignment is 13.77
- Which would you choose based on this information?

Utilities Data

- I then applied k-means with $K = 3$ using all 8 of the variables (standardized)
- I ran it several times using randomly generated cluster means $\{m_k : k = 1, \ldots, K\}$
- I chose the cluster assignment that gave the smallest $W(C)$

Utilities Data

- The cluster assignments are:

  1. Arizona, Central, Florida, Kentucky, Oklahoma, Southern, Texas
  2. Boston, NY, Hawaiian, New England, Pacific, San Diego, United
  3. Commonwealth, Idaho, Madison, Nevada, Northern, Puget, Wisconsin, Virginia