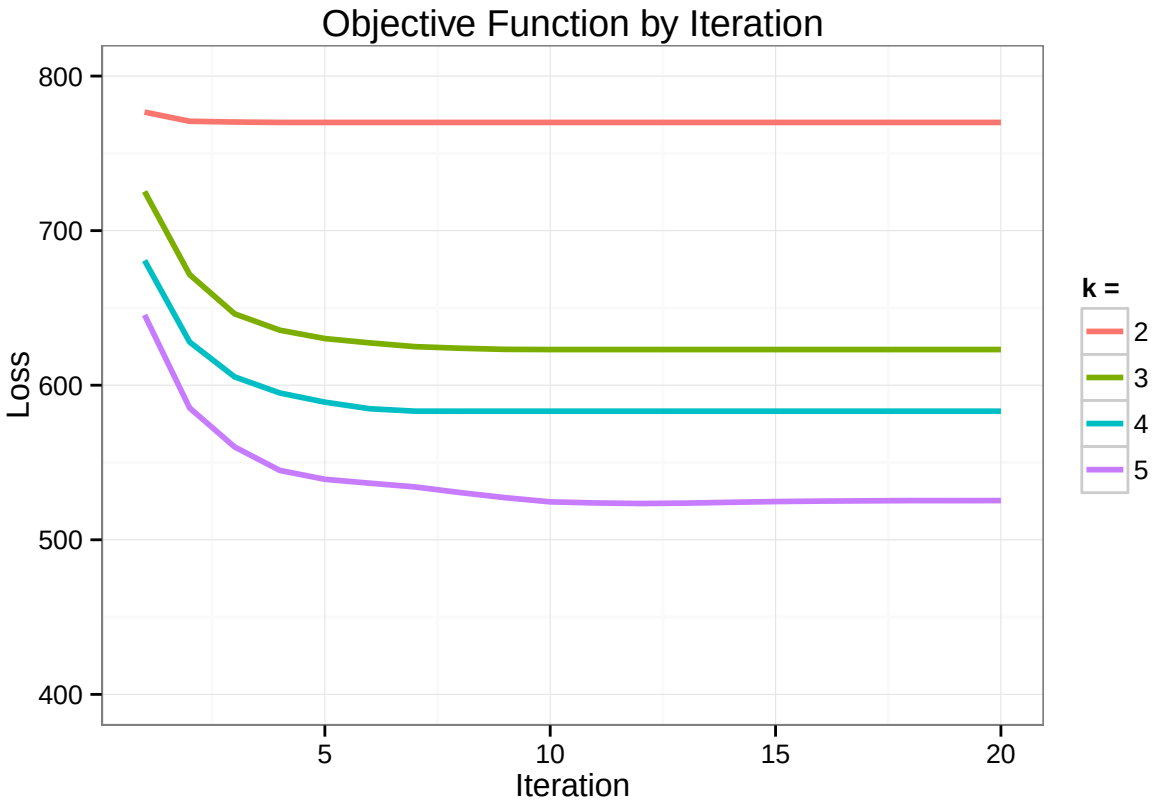
COMS 4721 - HW4

Jeff Hudson (jdh2182)

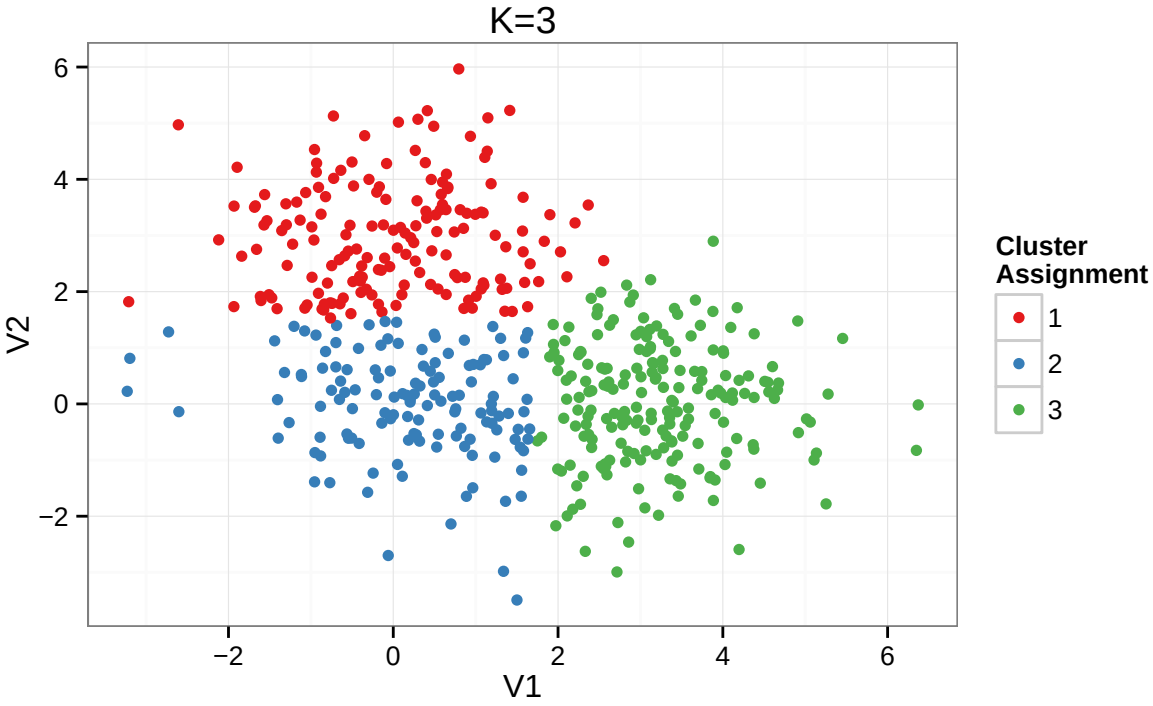
Tuesday, April 14, 2015

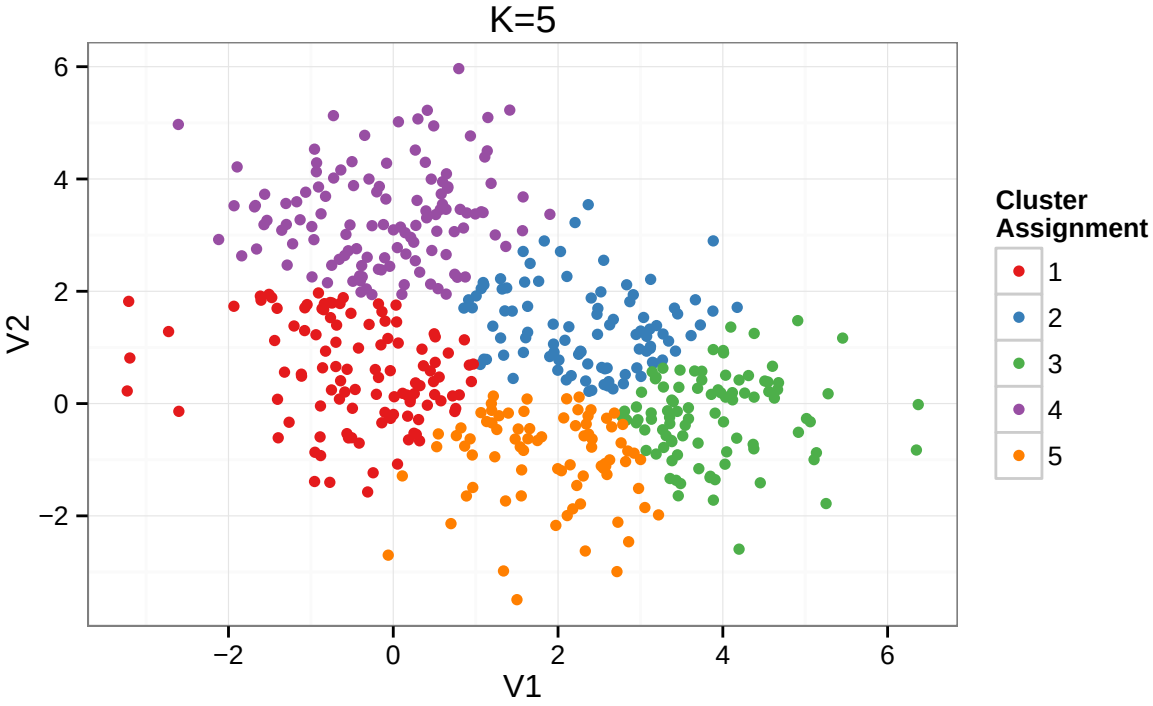
Problem 1

1. For $K = 2, 3, 4, 5$, plot the value of the K-means objective function per iteration for 20 iterations (the algorithm may converge before that).



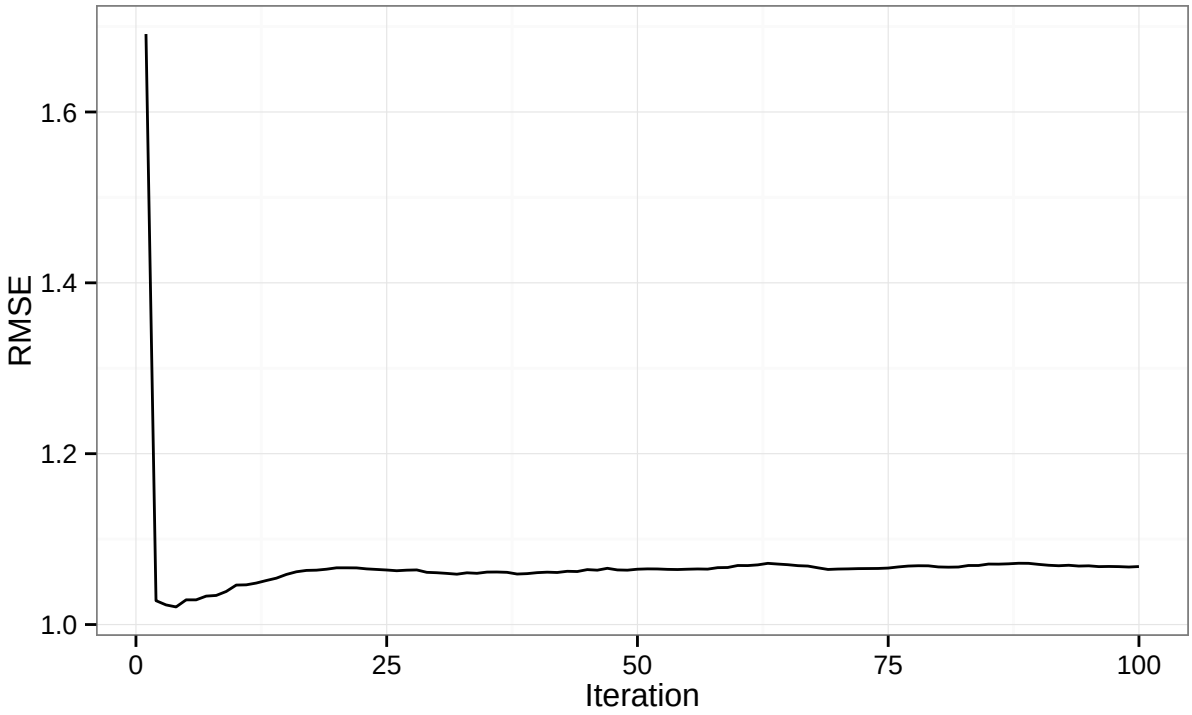
2. For $K = 3, 5$, plot the 500 data points and indicate the cluster of each for the final iteration by marking it with a color or a symbol.

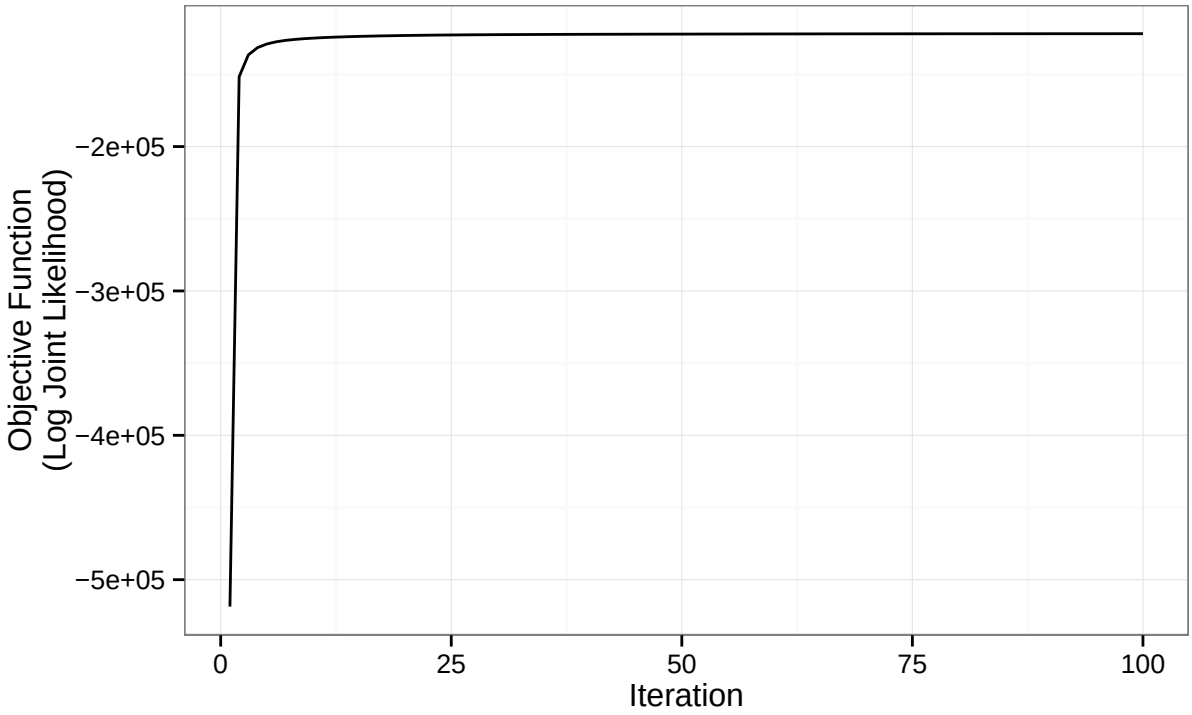




Problem 2

1. Plot the RMSE of your predictions on this test set as a function of training iteration. On a separate plot show the log joint likelihood as a function of iteration.





2. Pick three movies and for each movie find the 5 closest movies according to Euclidean distance using their respective locations v_i. List the query movie, the five nearest movies and their distances. A mapping from index to movie is given with the data.

```
## [1] "Five movies closest to: Pretty Woman (1990)"
```

```
##                           movie_name    dist
## 692  American President, The (1995) 0.82318
## 402                    Ghost (1990) 1.05912
## 393           Mrs. Doubtfire (1993) 1.16184
## 216  When Harry Met Sally... (1989) 1.16465
## 1042              Just Cause (1995) 1.18076
```

```
## [1] "Five movies closest to: Star Trek: The Wrath of Khan (1982)"
```

```
##                                       movie_name    dist
## 227 Star Trek VI: The Undiscovered Country (1991) 1.07117
## 230          Star Trek IV: The Voyage Home (1986) 1.08111
## 89                             Blade Runner (1982) 1.29100
## 229    Star Trek III: The Search for Spock (1984) 1.31910
## 164                            Abyss, The (1989) 1.40655
```

```
## [1] "Five movies closest to: Lion King, The (1994)"
```

```
##                                    movie_name    dist
## 393                      Mrs. Doubtfire (1993) 1.07006
## 95                              Aladdin (1992) 1.07110
## 1035                     Cool Runnings (1993) 1.07745
## 588                Beauty and the Beast (1991) 1.09668
## 435  Butch Cassidy and the Sundance Kid (1969) 1.13855
```

3. Perform K-means on the $u_1, \ldots, u_{N1}$ learned by your algorithm. Set $K = 30$. The centroids can be interpreted as personality types (as far as movies are concerned). Pick 5 centroids. For each centroid, characterize the cluster by showing the 10 movies with the largest (most positive) dot product to that centroid.

```
## [1] "Centroid #1"
##       [,1]                                    [,2]
##  [1,] "My Fair Lady (1964)"                   "4.84462"
##  [2,] "Strictly Ballroom (1992)"              "4.73526"
##  [3,] "Babe (1995)"                           "4.65857"
##  [4,] "Philadelphia Story, The (1940)"        "4.57748"
##  [5,] "In & Out (1997)"                       "4.57549"
##  [6,] "Lion King, The (1994)"                 "4.57245"
##  [7,] "Searching for Bobby Fischer (1993)"    "4.52688"
##  [8,] "Persuasion (1995)"                     "4.52212"
##  [9,] "Close Shave, A (1995)"                 "4.51855"
## [10,] "Local Hero (1983)"                     "4.51476"
```

```
## [1] "Centroid #2"
##       [,1]                                  [,2]
##  [1,] "Star Wars (1977)"                    "4.59404"
##  [2,] "Shawshank Redemption, The (1994)"    "4.55189"
##  [3,] "Pulp Fiction (1994)"                 "4.51885"
##  [4,] "Game, The (1997)"                    "4.47477"
##  [5,] "Empire Strikes Back, The (1980)"     "4.41219"
##  [6,] "Princess Bride, The (1987)"          "4.40865"
##  [7,] "Return of the Jedi (1983)"           "4.35351"
```

```
##  [8,] "Usual Suspects, The (1995)"       "4.33941"
##  [9,] "Titanic (1997)"                   "4.32144"
## [10,] "Great Escape, The (1963)"         "4.29109"
```

```
## [1] "Centroid #3"
##       [,1]                                      [,2]
##  [1,] "Godfather, The (1972)"                   "4.79991"
##  [2,] "Fargo (1996)"                            "4.77996"
##  [3,] "L.A. Confidential (1997)"                "4.66361"
##  [4,] "One Flew Over the Cuckoo's Nest (1975)" "4.66035"
##  [5,] "Shawshank Redemption, The (1994)"        "4.6242"
##  [6,] "Lawrence of Arabia (1962)"               "4.61588"
##  [7,] "Boot, Das (1981)"                        "4.60877"
##  [8,] "Star Wars (1977)"                        "4.58825"
##  [9,] "Silence of the Lambs, The (1991)"        "4.55614"
## [10,] "Pulp Fiction (1994)"                     "4.52036"
```

```
## [1] "Centroid #7"
##       [,1]                                [,2]
##  [1,] "Full Monty, The (1997)"            "5.1186"
##  [2,] "Lost Highway (1997)"               "4.90427"
##  [3,] "Leaving Las Vegas (1995)"          "4.86438"
##  [4,] "Close Shave, A (1995)"             "4.60536"
##  [5,] "L.A. Confidential (1997)"          "4.59242"
##  [6,] "Trainspotting (1996)"              "4.55942"
##  [7,] "City of Lost Children, The (1995)" "4.48954"
##  [8,] "Wrong Trousers, The (1993)"        "4.42577"
##  [9,] "Princess Bride, The (1987)"        "4.37657"
## [10,] "Star Wars (1977)"                  "4.35942"
```

```
## [1] "Centroid #10"
##       [,1]                                                    [,2]
##  [1,] "Apt Pupil (1998)"                                      "4.91418"
##  [2,] "Henry V (1989)"                                        "4.69884"
##  [3,] "As Good As It Gets (1997)"                             "4.67135"
##  [4,] "Mrs. Brown (Her Majesty, Mrs. Brown) (1997)"           "4.60738"
##  [5,] "Richard III (1995)"                                    "4.57819"
##  [6,] "Wallace & Gromit: The Best of Aardman Animation (1996)" "4.57137"
##  [7,] "Cold Comfort Farm (1995)"                              "4.5621"
##  [8,] "Jackie Brown (1997)"                                   "4.55981"
##  [9,] "Good Will Hunting (1997)"                              "4.5398"
## [10,] "Antonia's Line (1995)"                                 "4.48312"
```