11-755: MLSP

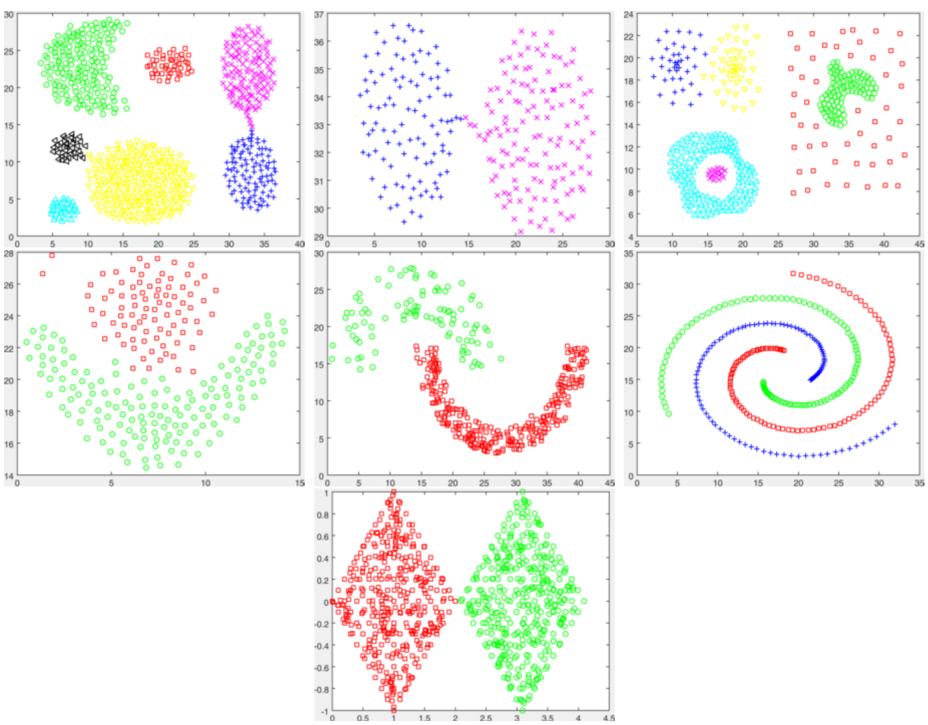
HW #3: Clustering and EM

Nikolas Wolfe

December 15, 2015

Problem 1: K-Means & Spectral Clustering

You are given a number of toy datasets. The visulized groud truth clustering configuration of these data: Aggregation (K=7), Bridge (K=2), Compound (K=6), Flame (K=2), Jain (K=2), Spiral (K=3) and TwoDiamond (K=2) are shown as follows:



Answer:

I implemented this code in Java, using the Efficient Java Matrix Library (EJML) package to do matrix multiplication and Eigen/Singular Value decomposition. In general, spectral clustering works a lot better than K-means and is able to determine a lot of structure (such as the spiral pattern) that k-means is unable to find.

In terms of evaluation, I ran the clustering algorithms and generated the confusion matrix between the most common label that emerges in each cluster and their actual labels. A diagonal matrix in this case is of course perfect performance. The accuracy can be determined from looking at the counts of labeled points for each class.

Dataset	Optimal K	Group Counts
Aggregation	7	$[K_1, K_2, K_3, K_4, K_5, K_6, K_7] = [45, 170, 102, 273, 34, 130, 34]$
Compound	6	$[K_1, K_2, K_3, K_4, K_6] = [50, 92, 38, 45, 158, 16]$
Spiral	3	$[K_1, K_2, K_3] = [101, 105, 106]$
Bridge	2	$[K_1, K_2] = [102, 13]$
Flame	2	$[K_1, K_2] = [87, 153]$
Jain	2	$[K_1, K_2] = [276, 97]$
Two Diamonds	2	$[K_1, K_2] = [400, 400]$

With this information, we can directly evaluate the quality of the Spectral and K-means clustering algorithms by simply comparing the diagonal of their confusion matrices (most common group label vs. actual label) to the vectors listed above above. To run the java package, simply run the following from the root of this project:

```
java -jar nwolfe_hw3.jar
```

This will run the main method in the class `ClusterDriver.java`, which will then iterate through a small range of σ values which are known to produce generally good spectral clusters, though it is often the case that several random restarts are required in order for the spectral clustering algorithm to settle into the optimal cluster groupings. In this report the best clusters have been chosen, though it is the case that when *you* run it, it might produce an optimal grouping.

This code also produces the results as expected by the provided MATLAB scripts as CSV files in the root directory in which the jar is run. To generate the visualizations of the data, you may select the individual files you want and use the `visualize.m` script as provided. Alternatively, you may use the results I have included in the report and simply go to the `external` folder and run the `gen_all_figs.m` script.

```
cd external/
matlab gen_all_figs.m
```

Alternatively, you could just look at the pictures in this here report. That's probably easier.

Questions

In your report, answer the following questions:

1. Is the objective function of k-means a convex one?

   Answer: The objective function which we are trying to minimize is the sum of Euclidean distances from each cluster centroid to each cluster datapoint. However, the number of clusters K is discrete, which means we cannot take the derivative of this function and determine the optimal grouping (i.e. the minimum distance sum over all clusters) either analytically or through gradient descent or some other minimization technique. (This is also the case for spectral clustering.) Furthermore, the 2nd derivative of the Euclidean distance function is not zero. Because there is no way to find the optimal grouping (and perhaps no grouping actually exists), this is not a convex optimization problem.

2. Can the iterative updating scheme in k-means return a solution which achieves the global minimum of the k-means objective function?

   Answer: Newp. No. Sorry! As stated above, there is no guaranteed minimum for the objective function because it isn't convex. There may be an infinite number of viable ways to cluster the data.

3. If yes, why? If not, why not and are there any practical ways to partially fix this?

   Answer: One of the ways to fix this is, first off, to attempt to pick the initialization of the cluster centroids according to some logic. In this case we happen to know the optimal value of K from the labeled data. This may result in a better teasing apart of the datapoints and less moving around before the algorithm converges. Another way might be to use bottom up or top down clustering, i.e. to start with a single cluster containing all the data, perturb the centroid and then continually break the clusters apart until an optimal value of K is reached. Or we could start with a large number of clusters and then continually merge them according to some criteria until we reach a desired value of K.

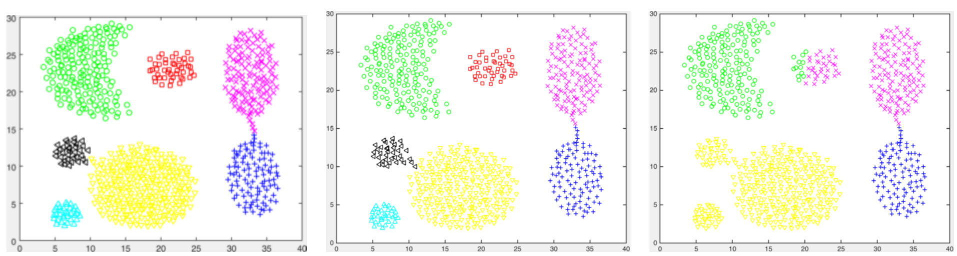
Results

For each problem I have printed the output of the algorithm after a single run. At the beginning of each section is a comparison of 3 pictures.

Aggregation

From left to right, the images are:

1. Ground Truth
2. Spectral Clustering Result
3. K-Means Clustering Result



Aggregation: Spectral Clustering

```
Reading data values in data/Aggregation.csv...
Reading data values in data/Bridge.csv...
Reading data values in data/Compound.csv...
Reading data values in data/Flame.csv...
Reading data values in data/Jain.csv...
Reading data values in data/Spiral.csv...
Reading data values in data/TwoDiamonds.csv...

=============== AGGREGATION: SPECTRAL CLUSTERING ===============
With sigma = 0.47
```

```
label  1    2    3    4    5    6    7
------------------------------------------------------------
1 |    45   0    0    0    0    0    0
2 |    0    170  0    0    0    0    0
3 |    0    0    1    0    0    0    0
4 |    0    0    0    270  0    0    0
5 |    0    0    0    0    34   0    0
6 |    0    0    101  0    0    130  0
7 |    0    0    0    3    0    0    34

With sigma = 0.48

label  1    2    3    4    5    6    7
------------------------------------------------------------
1 |    45   0    0    0    0    0    0
2 |    0    170  0    0    0    0    0
3 |    0    0    102  0    0    2    0
4 |    0    0    0    270  0    0    0
5 |    0    0    0    0    34   0    0
6 |    0    0    0    0    0    128  0
7 |    0    0    0    3    0    0    34

With sigma = 0.49

... Output truncated ...
```

In the second run above, the diagonal is almost perfect, save for a few items misplaced in groups 3 and 7.

Aggregation: K-Means Clustering

```
===================== AGGREGATION: K MEANS =====================

label  1    2    3    4    5    6    7
------------------------------------------------------------
1 |    0    0    0    0    0    0    0
2 |    17   170  0    0    0    0    0
3 |    0    0    102  0    0    3    0
4 |    0    0    0    273  34   0    34
5 |    0    0    0    0    0    0    0
6 |    28   0    0    0    0    127  0
```

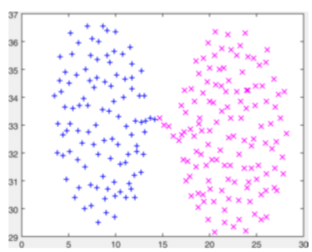
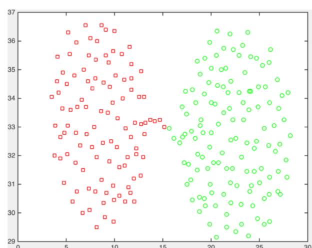
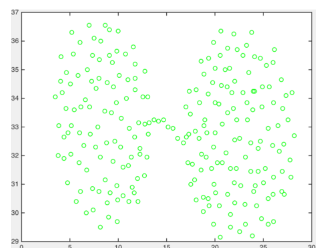
```
7 |     0     0     0     0     0     0     0

-----------------------------------------------------------
```

In the k-means clustering it's clear that several groups did not even emerge from the data (though this does not mean they were empty.)

Bridge

From left to right, the images are:

1. Ground Truth
2. Spectral Clustering Result
3. K-Means Clustering Result

Bridge: Spectral Clustering

```
=============== BRIDGE: SPECTRAL CLUSTERING ===============
With sigma = 0.42

... Output truncated ...

With sigma = 0.44

label    1    2
-------------------
1 |      102  2
2 |      0    128

With sigma = 0.45
```

```
label     1    2
-------------------
1 |       58   58
2 |       44   72
```

In one case the grouping was nearly perfect, and in another the data was more or less spread around a different kind of grouping.

Bridge: K-Means Clustering

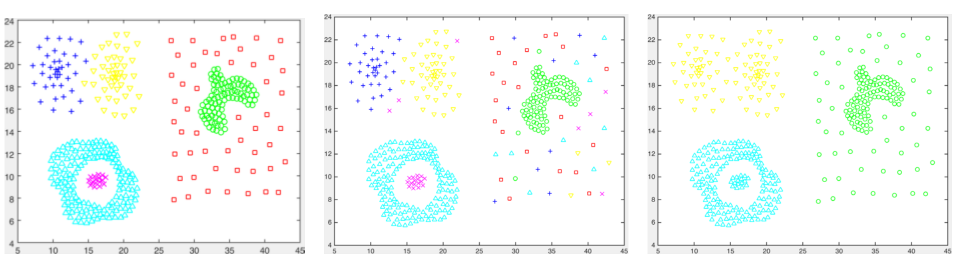
```
===================== BRIDGE: K MEANS =====================

label     1    2
-------------------
1 |       0    0
2 |       102  130

----------------------------------------------------------
```

Again, K-means was not able to find the right clustering...

Compound

From left to right, the images are:

1. Ground Truth
2. Spectral Clustering Result
3. K-Means Clustering Result



Compound: Spectral Clustering

```
=============== COMPOUND: SPECTRAL CLUSTERING ===============

With sigma = 0.154

label  1    2    3    4    5    6

---------------------------------------------------
1 |    21   0    1    1    0    0
2 |    6    92   0    0    0    0
3 |    12   0    33   0    0    0
4 |    3    0    1    44   0    0
5 |    5    0    2    0    158  0
6 |    3    0    1    0    0    16

With sigma = 0.155

label  1    2    3    4    5    6

---------------------------------------------------
1 |    20   0    2    1    0    0
2 |    6    92   1    44   0    0
3 |    11   0    35   0    0    0
4 |    0    0    0    0    0    0
5 |    7    0    0    0    158  0
6 |    6    0    0    0    0    16

... Output truncated ...
```

This grouping is particularly interesting because the groups in the compound picture were very difficult to tease apart, and yet we see here a diagonal clearly taking shape, except for some overlap with group 1.

Compound: K-Means Clustering

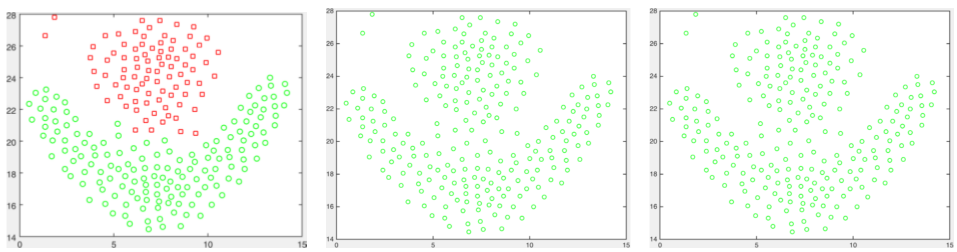
```
===================== COMPOUND: K MEANS =====================

label  1    2    3    4    5    6

---------------------------------------------------
1 |    0    0    0    0    0    0
2 |    48   92   0    0    0    0
3 |    0    0    38   1    0    0
```

```
4 |     2    0    0    44   0    0
5 |     0    0    0    0    158  16
6 |     0    0    0    0    0    0

-----------------------------------------------------------
```

Flame

From left to right, the images are:

1. Ground Truth
2. Spectral Clustering Result
3. K-Means Clustering Result



Aggregation: Flame Clustering

```
=============== FLAME: SPECTRAL CLUSTERING ===============
With sigma = 0.735

label    1    2

-------------------
1 |      0    0
2 |      87   153
```

Flame: K-Means Clustering

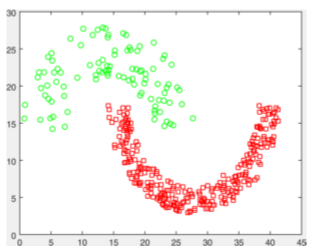
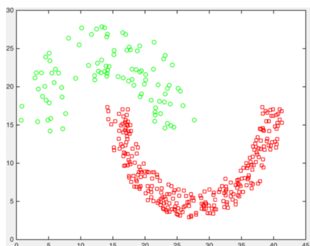
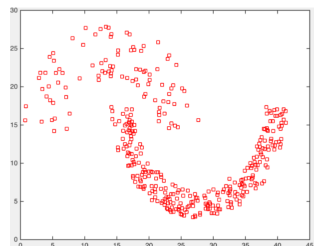
```
===================== FLAME: K MEANS =====================

label    1    2

-------------------
1 |      0    0
```

```
2 |      87   153

----------------------------------------------------------
```

Jain

From left to right, the images are:

1. Ground Truth
2. Spectral Clustering Result
3. K-Means Clustering Result

Jain: Spectral Clustering

```
=============== JAIN: SPECTRAL CLUSTERING ===============

With sigma = 0.303

label    1    2
-------------------
1 |      276  0
2 |      0    97

With sigma = 0.304

label    1    2
-------------------
1 |      276  97
2 |      0    0

... Output truncated ...
```

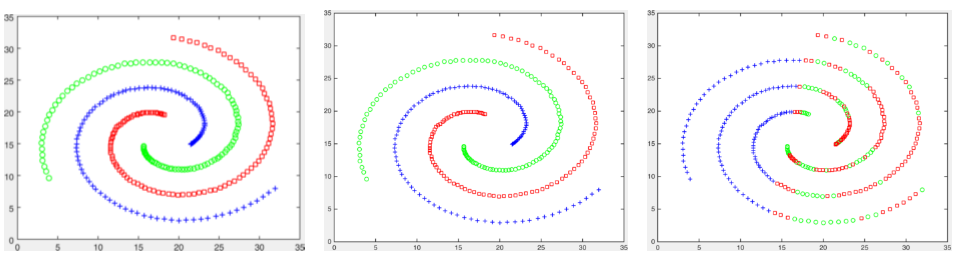
Jain: K-Means Clustering

```
===================== JAIN: K MEANS =====================

label    1    2
-------------------
1 |      276  97
2 |      0    0

----------------------------------------------------------
```

Spiral

From left to right, the images are:

1. Ground Truth
2. Spectral Clustering Result
3. K-Means Clustering Result



Spiral: Spectral Clustering

```
=============== SPIRAL: SPECTRAL CLUSTERING ===============
With sigma = 0.2

label  1    2    3
---------------------------
1 |    101  0    0
2 |    0    105  0
3 |    0    0    106
```

```
With sigma = 0.30000000000000004

label  1    2    3
---------------------------
1 |    101  0    0
2 |    0    105  0
3 |    0    0    106

With sigma = 0.4

label  1    2    3
---------------------------
1 |    101  0    0
2 |    0    105  0
3 |    0    0    106

... Output truncated ...
```

Given the shape in this case we can really see the difference between spectral clustering and k-means. This grouping was found perfectly, all zeros in the off-diagonal elements, with even a variable number of settings for σ. Looking at the k-means result below, it's clear that there's a real distinction here.

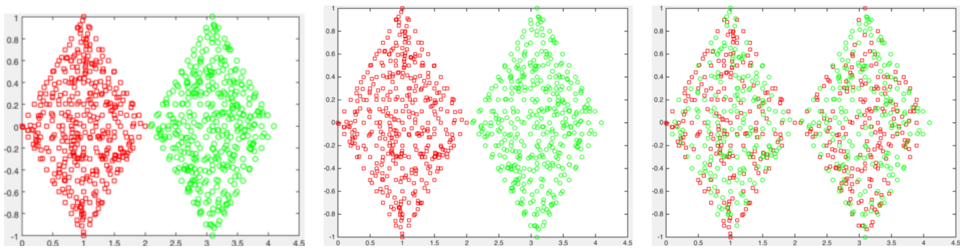
Spiral: K-Means Clustering

```
===================== SPIRAL: K MEANS =====================

label  1    2    3
---------------------------
1 |    66   54   53
2 |    0    0    0
3 |    35   51   53

-----------------------------------------------------------
```

Two Diamonds

From left to right, the images are:

1. Ground Truth
2. Spectral Clustering Result
3. K-Means Clustering Result



In the above case it appears that the two centroids in the k-means cluster were positioned in such a way that the points are ambiguously scattered, although I should note that I did not control for bad k-means results... I have only tried to present the best spectral clustering results because they're more fun anyway.

Two Diamonds: Spectral Clustering

```
=============== TWO DIAMONDS: SPECTRAL CLUSTERING ===============
With sigma = 20.0

label    1    2
-------------------
1 |      400  0
2 |      0    400

With sigma = 21.0

label    1    2
-------------------
1 |      204  198
2 |      196  202

... Output truncated ...
```

Two Diamonds: K-Means Clustering

```
==================== TWO DIAMONDS: K MEANS ====================

label    1    2
------------------
1 |      203  196
2 |      197  204

-------------------------------------------------------
```

Questions & Conclusion

Compare your spectral clustering results with k-means. It is natural that on certain hard toy examples, both method won't generate perfect results? In your report, briefly analyze what is the advantage or disadvantage of spectral clustering over k-means. Why it is the case?

Answer:

It is natural in the sense that the Gaussian kernel function used in spectral clustering is good at separating certain kinds of data and not others. Clearly separable groups, even if they have complicated boundaries (such as the spiral) are

Overall, I found that SVD was better at finding the basis vectors for the projection of the points into k dimensional space. Furthermore in many cases it required a lot of random restarts before the groups coalesced into something reasonable.

Problem 2: Expectation Maximization

Let Z be the sum of two random variables X and Y, ($Z = X + Y$), where X and Y are drawn independently from two discrete probability distributions with probability mass functions defined as:

$$P(X = n) = p_1(1 - p_1)^{n-1} \tag{1}$$

$$P(Y = n) = p_2(1 - p_2)^{n-1} \tag{2}$$

Given samples of Z, derive an EM algorithm to estimate p_1 and p_2.

Answer: We have two discrete probability distributions. Samples of Z are always the sum of X and Y, so Z is a multinomial distribution. We want to use the EM algorithm to estimate the maximum likelihood values of p_1 and p_2 that generated the samples of Z. We don't know p_1 and p_2 from observing Z, so we have to iteratively *fractionate* the observations and use them to re-estimate our expectation for the values of p_1 and p_2. because $p1$ and $p2$ are the only parameters of the above distributions, we can alternatively write them as:

$$P(X) = p_1(n) \tag{3}$$
$$P(Y) = p_2(n) \tag{4}$$

We need a way to calculate the parameters $p1$ and $p2$, however. We must estimate this from the data. The total log probabilities of X and Y are the sum of the probabilities of each value of n times their counts c_n. This is:

$$log(P(X)) = \sum_n c_n \; log(p_1) + log(1 - p_1)(n - 1) \tag{5}$$
$$log(P(Y)) = \sum_n c_n \; log(p_2) + log(1 - p_2)(n - 1) \tag{6}$$

To find the maximum likelihood values of p_1 and p_2 we differentiate this equation w.r.t. P and set it to zero:

$$\frac{\partial}{\partial p_1} log(P(X)) = \frac{\partial}{\partial p_1}\left(\sum_n c_n \; log(p_1) + log(1 - p_1)(n - 1)\right) \tag{7}$$
$$0 = \frac{\partial}{\partial p_1}\left(\sum_n c_n \; log(p_1) + log(1 - p_1)(n - 1)\right) \tag{8}$$
$$0 = \sum_n \frac{c_n}{p_1} - \sum_n \frac{n-1}{(1 - p_1)} \tag{9}$$
$$\sum_n \frac{c_n}{p_1} = \sum_n \frac{n-1}{(1 - p_1)} \tag{10}$$
$$p_1 = \sum_n \frac{c_n}{c_n + n - 1} \tag{11}$$

Likewise for $p_2 = \sum_n \frac{c_n}{c_n+n-1}$. This works for the non-log form below because the value that maximizes the logarithmic from will also maximize the original. This of course assumes we know the value of c_n which depends on the value of our hidden variables. So we must estimate it. Over all observations O, we'll say the number of observations is given by $size(O)$. Here is the formula for the count of a given value n which forms part of the sum of the observation o:

$$c_n = \sum_{o \in O} \; size(O) \; p_{1,2}(n, o - n|o) \tag{12}$$

p can be either p_1 or p_2 based on our current estimates for them. Now that we have this, our algorithm is as follows. H represents our hidden variables, and O is the observation.

1. **Initialize** $P(H)$, $P(O|H)$
   (a) $P(H)$ in our case is $P(X)$ and $P(Y)$, which we can alternatively write as $P(X, Z-X)$ because $Z = X + Y$. Randomly initialize these values as well.
   (b) $P(O|H)$ is initially a guess for $P(Z|X, Z-X)$. We can randomly initialize these values for all Z in our observation.
2. **Estimate** $P(H|O)$ for each H ($p1$ and $p2$), for each called out number
   (a) This is a value we must estimate over all the possible combinations of X and Y which sum to Z. If $X = n$ and $Y = Z - n$, in general we can calculate this as follows:
   $$P(n, Z-n|Z) = \frac{p_1(n)\ p_2(Z-n)}{\sum_n\ p_1(n)\ p_2(Z-n)} \tag{13}$$
   (b) This summation is over all discrete values of n which can contribute to the value of Z, and we know $p1$, $p2$ from (3) and (4) and (11) and (12).
3. **Fractionate** O with each value of H, with weight $P(H|O)$ as calculated in the previous step. In this case the value of the combination Y and X will have the same weight because any value of X determines Y. But the fractions of the multiplied values will be different.
4. **Re-estimate** $P(O|H)$ for every value of O and H. This is calculated as follows, where c is the count for discrete value of observation O:
   $$P(O|H) = \frac{c\ P(H|O)}{\sum_O c\ P(H|O)} \tag{14}$$
   In our case, this becomes:
   $$P(Z|X=n, Z-n) = \frac{c_n\ P(n, Z-n|Z)}{\sum_n c_n\ P(n, Z-n|Z)} \tag{15}$$
   We of course have c_n from (12).

5. **Re-estimate** $P(H)$

$$P(H) = \frac{\sum_O c\ P(H|O)}{\sum_{H'} \sum_O c\ P(H'|O)} \tag{16}$$

In our case, this becomes:

$$P(p1, p2) = \frac{\sum_n c_n\ P(n, Z-n|Z)}{\sum_{P'} \sum_n c_n\ P'(n, Z-n|Z)} \tag{17}$$

Where P' represents our current model parameter estimates for p_1 and p_2, the Maximum Likelihood values of which are derived above in (11) and (12).

6. If not converged, **return** to 2

Got it? The answer is 42.