# Unsupervised Learning of Cancer Data

By Connor Leipelt

Department of Mathematics and Statistics

University of Massachusetts Amherst

# Introduction

- Goal: Investigate different unsupervised learning techniques, specifically clustering types, on cancerous tumor data.
- Reasoning: Cancer growth is notoriously difficult to generalize to large populations. Through clustering techniques one can research the correlation in each cluster group to find an explanation for specific growth patterns.
- Types of Clustering Methods being used:
  - Gaussian Mixture Model with EM algorithm
  - Hierarchical Agglomerative Clustering
  - K means Clustering

# Data Background

- The data for this project came from Professor Wang at Union College. I studied the growth rates of cancerous tumors on mice who underwent DMPA/TPA treatment and were given Imiquimod or 5-FU for chemotherapy.
- There were 45 mice being investigated with all mice having 3-4 tumors (137 total tumors). There are two main data sets, one of all tumors and one of mice.
- The mice received chemotherapy after a tumor reached a diameter of 3-4 mm allowing for two different types of growth rates to be calculated; initial growth rate and the growth rate after treatment.
- There were two doses of Imiquimod and 5-FU and a control group so 5 clusters is expected to look correct. Clusters have been tested for k = 2,3,4,5.

# Growth Rate Calculations

- We will be looking at the average annual growth rate (g) for each tumor. This rate can be calculated as follows: $g := (V_n / V_0)^{1/n} - 1$ where $V_0$ := initial volume, $V_n$ := volume at day $n$.
- Assuming that the next day's volume is dependent on the growth of the previous at the same rate for each day we can prove the growth rate.

Proof:

Define $V_1 = V_0 + g \times V_0 = V_0(1 + g)$ and $V_2 = V_1 + g \times V_1 = V_1(1 + g) = V_0(1 + g)^2$

Inductively, we see that $V_n = V_0(1 + g)^n$ which gives $(V_n / V_0)^{1/n} - 1 = g$ concluding the proof.

∎

- Since we now have an initial and treatment growth rate, we can start the cluster methods.

# Gaussian Mixed Model & EM Algorithm

- This method creates a generative model of our data.
- It creates a mixed Gaussian by combining Gaussians with specific weights labeled $\pi_k$:

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

- The goal of this method is to iteratively refine this mixed model to accurately fit the clustered data.
- Taking the EM algorithm repeatedly increases the log-likelihood, meaning our classification of the clusters are converging!
- Note that this method can get stuck at local optima so initialization of the parameters is important.

- E-Step: - start with chosen $\mu_k$, $\Sigma_k$, and $\pi_k$
  - Note that these can be randomly generated
  - For each data point $x_n$ we compute the responsibility for cluster $k$:

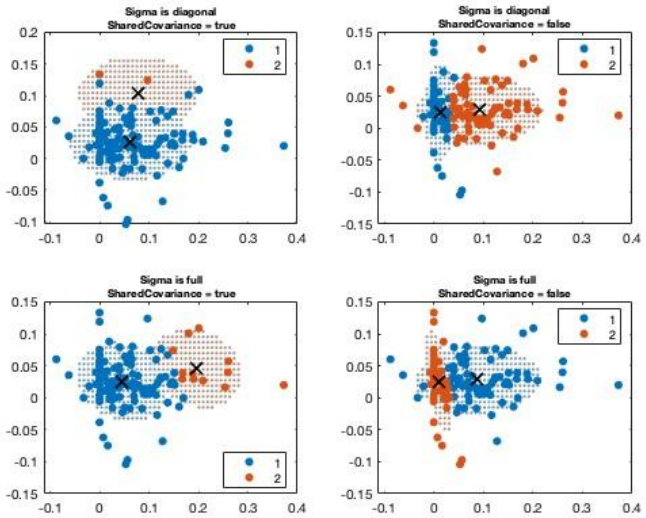
$$r_{nk} \triangleq p(z_n = k|\boldsymbol{x}_n, \boldsymbol{\theta}) = \frac{p(z_n = k|\boldsymbol{\theta}) p(\boldsymbol{x}_n|z_n = k, \boldsymbol{\theta})}{\sum_{k'=1}^{K} p(z_n = k'|\boldsymbol{\theta}) p(\boldsymbol{x}_n|z_n = k', \boldsymbol{\theta})}$$

- M-step: - Fix $r_{n,k}$ and update $\mu_k$, $\Sigma_k$, and $\pi_k$ as follows $m_k = \Sigma_n r_{n,k}$(sum over n), $\pi_k = (m_k/m)$ with $m = \Sigma_k m_k$(sum over k), $\mu_k = (1/m_k)\Sigma_n r_{n,k} x_n$(sum over n), $\Sigma_k = (1/m_k)\Sigma_n r_{n,k}(x_n - \mu_k)^T(x_n - \mu_k)$ (sum over n)

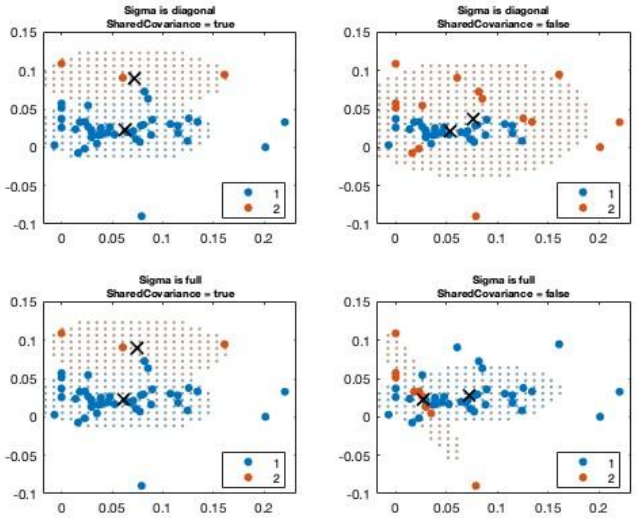
# Gaussian Mixed Model Findings

- Note that there is an importance on whether or not the covariance matrix is diagonal or full and if the covariance matrix is shared(shared => same covariance => same ellipsoid shape for all clusters). All combinations have been tried for k clusters, k = 2,3,4,5 (certain combinations result in ill-conditioned matrices so their plots are not present).

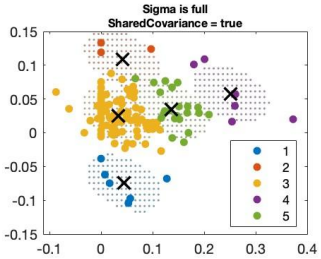
## Tumor Data Set k = 2



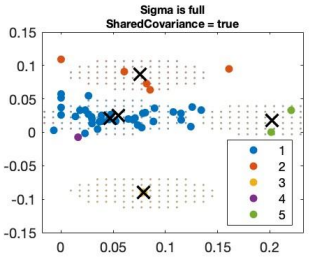

## Mice Data Set k = 2




## Tumor k =5




## Mice k=5




- Please take a look at all of the other graphs! :)

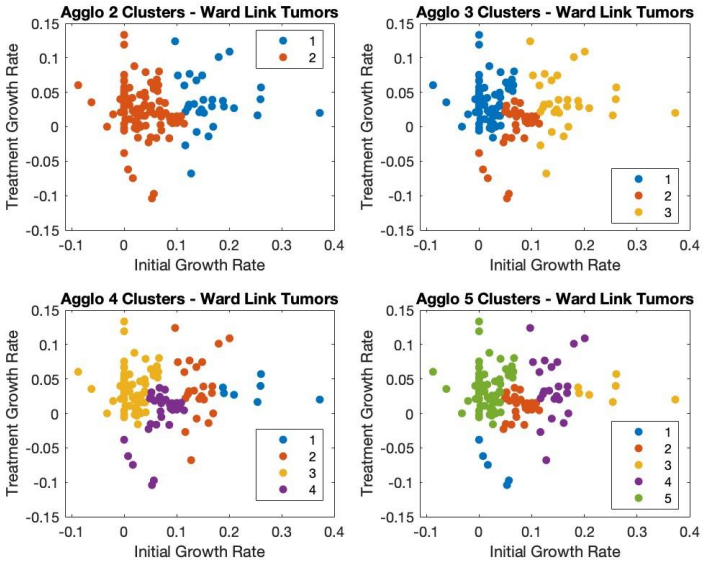
# Hierarchical Agglomerative Clustering

- Agglomerative clustering starts with making every data point being a cluster. At each stage certain clusters are grouped together based on the criterion specified. The process of this clustering can be seen through a dendrogram (read from the bottom up!). Four different methods have been implemented for this type of clustering.

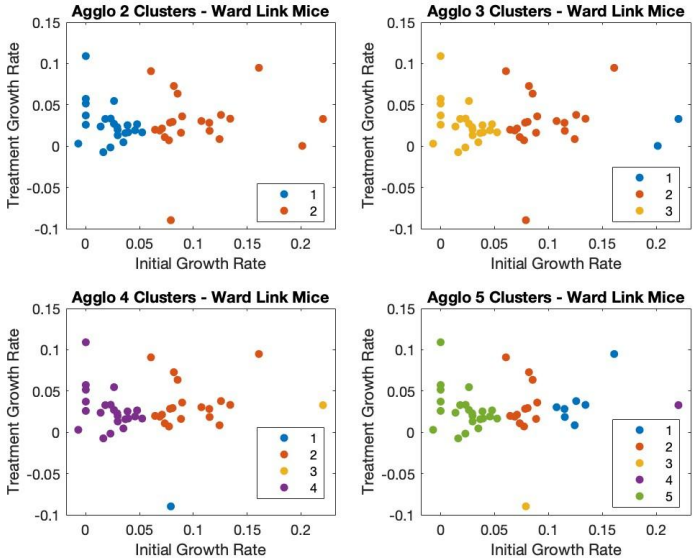
- Single Linkage:
  - At every stage combine the two closest clusters to form a new cluster.
- Average Linkage:
  - Combine the clusters that have the smallest average value over all points in both clusters. This means compute the distance between every point in A and every point in B and take the average of that value, whichever two clusters has the smallest average get combined.
- Complete Linkage:
  - Compute distance between all points in two clusters and take the maximum distance found, do this for all combinations of clusters and combine the two clusters with the smallest maximum distance.
- Ward Linkage
  - Combine two clusters and compute their within-cluster variance. Whichever two clusters create the smallest increase in covariance merge.

# Agglomerative Clustering Findings

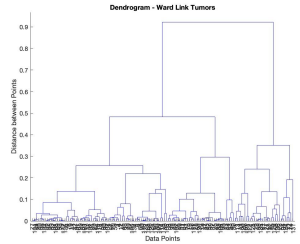
## Tumor Ward Cluster



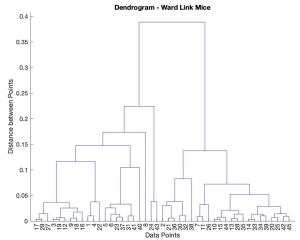

## Mice Ward Cluster




## Tumor Dendrogram Ward




## Mouse Dendrogram Ward



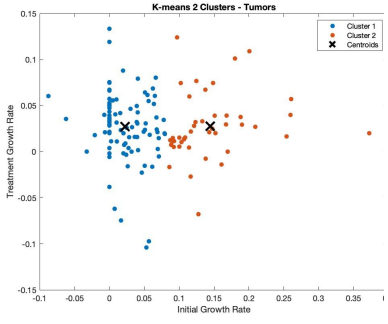

- There are 40 graphs for this model so please look at the extra graphs! :)

# K means Clustering

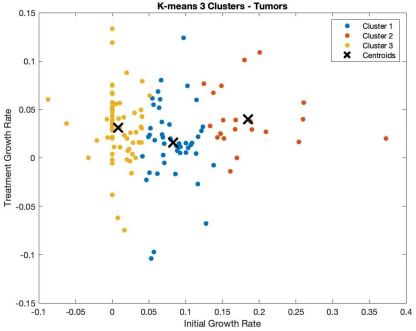
- K means clustering is a clustering method that depends on random starting points.
- First you specify how many clusters you want, k, then you randomly select k data points and make then the center of your clusters.
- From here you compute, for every data point, which cluster center it's closest to, and group the data points according to this metric.
- Now you take the average of all data points that are connected to one cluster and make that cluster's center the average.
- From here you rinse and repeat the method until the new center is the same as the previous for all clusters or you stop once the difference between the new and previous centers is less than some tolerance criterion, $\varepsilon$. Note that you will converge.

# K means Findings

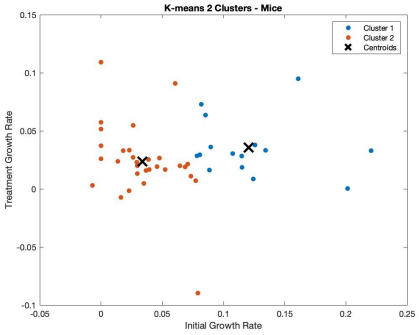
Tumor k = 2



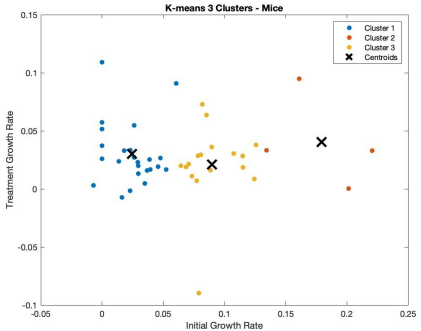

Tumor k =3



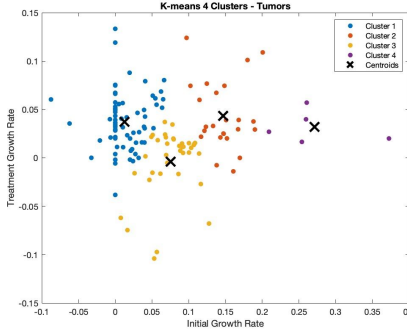

Mice k = 2



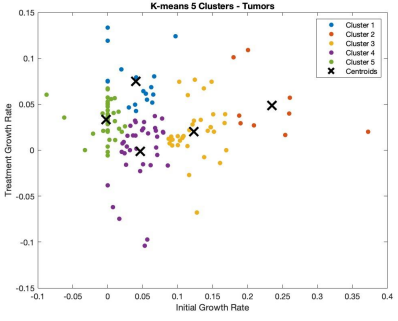

Mice k = 3



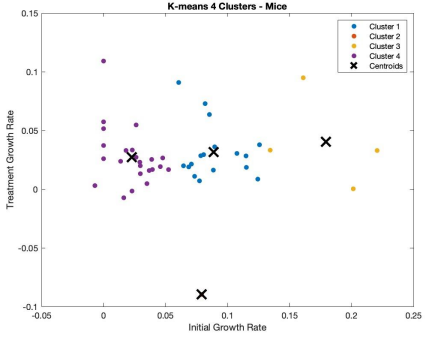

Tumor k = 4



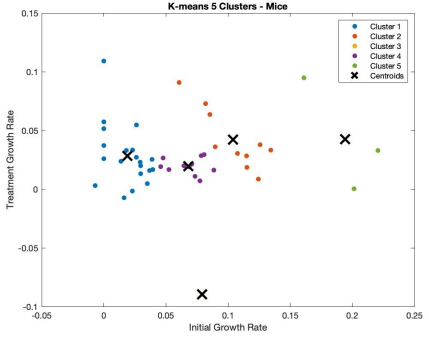

Tumor k = 5




Mice k = 4



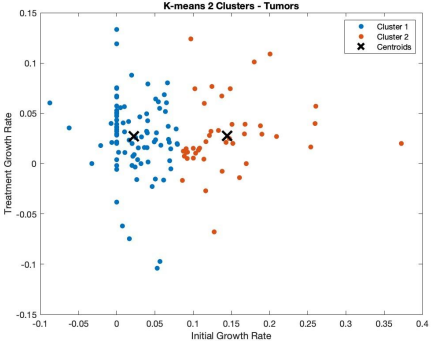

Mice k = 5



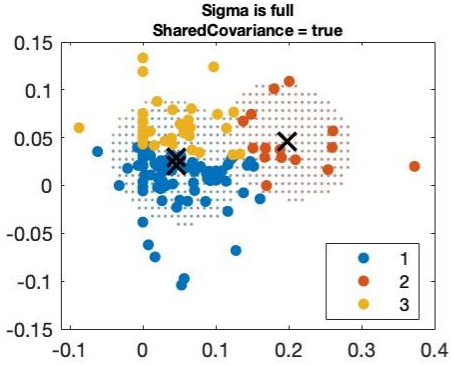
# Comparison and Conclusion (Tumors)

- We are investigating three clustering methods pertaining to all of the tumors: GMM, Agglo, and K means where GMM and Agglo have 4 different possible version.
- From qualitative evaluation we see that taking k = 5 seems to create fair clusterings, but we will investigate the highest performers for each k = 2,3,4,5. Note that we are looking for clusterings that seem realistic (mostly vertical/horizontal is not great, shows too much variance in one growth rate compared to the other).

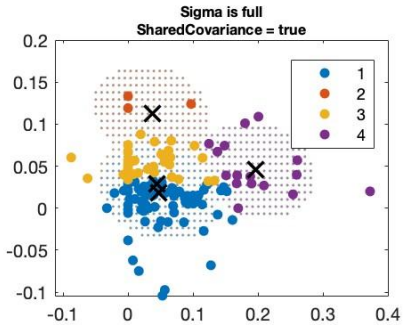
K = 2: K means



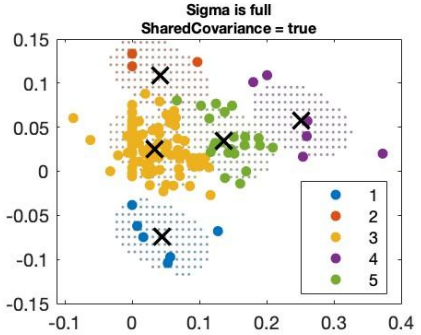

K = 3: GMM Full/True




K = 4: GMM Full/True



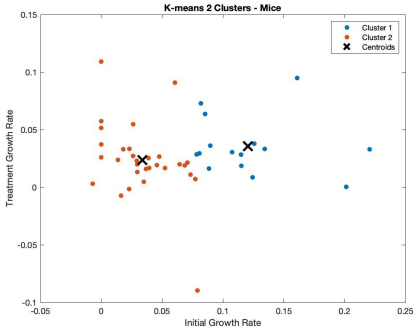

K = 5: GMM Full/True



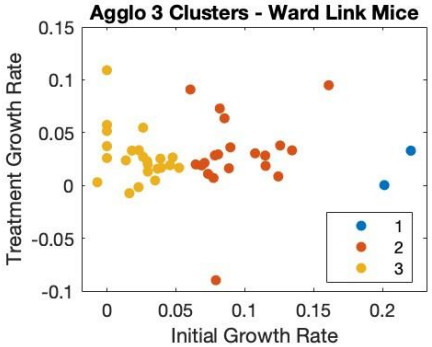
# Comparison and Conclusion (Mice)

- We are investigating three clustering methods pertaining to growth rates for each mouse: GMM, Agglo, and K means where GMM and Agglo have 4 different possible version.
- From qualitative evaluation we see that taking k = 5 seems to create fair clusterings, but we will investigate the highest performers for each k = 2,3,4,5. Note that we are looking for clusterings that seem realistic (mostly vertical/horizontal is not great, shows too much variance in one growth rate compared to the other).

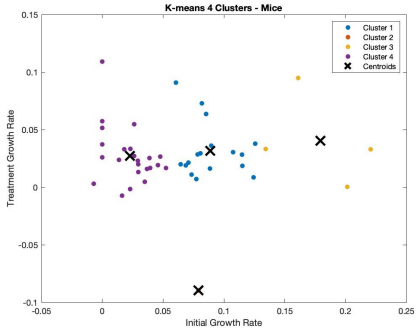
K = 2: K means



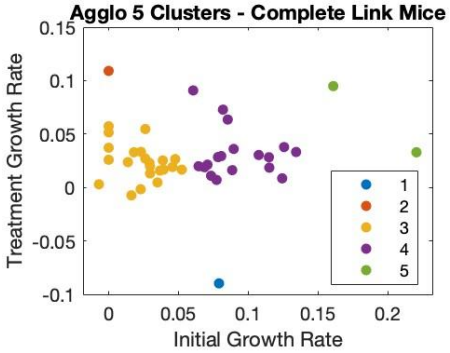

K = 3: Ward Link




K = 4: K means




K = 5: Complete Link



# Future Research Possibilities

- Investigate the specific clusters and look for correlated features other than growth rate.
- Look into processing speed of each algorithm to find the most efficient model.
- Run training vs test sets of data to validate how well each clustering method performs.
- Check to see if the 5 possible categories (Imiquimod-1, Imiquimod-2, 5-FU-1, 5-FU-2, and a control group) line up with the clusters.