An Investigation of Clustering: *k*-means and GMM

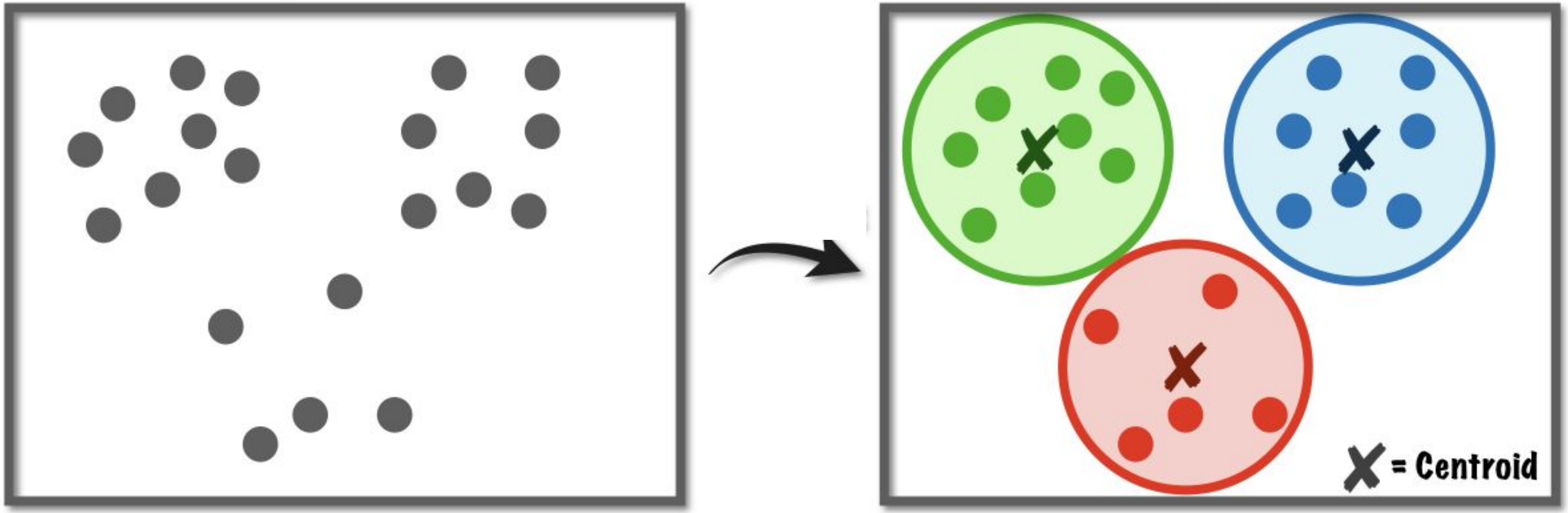
Emma Grossman

April 6, 2021

Clustering

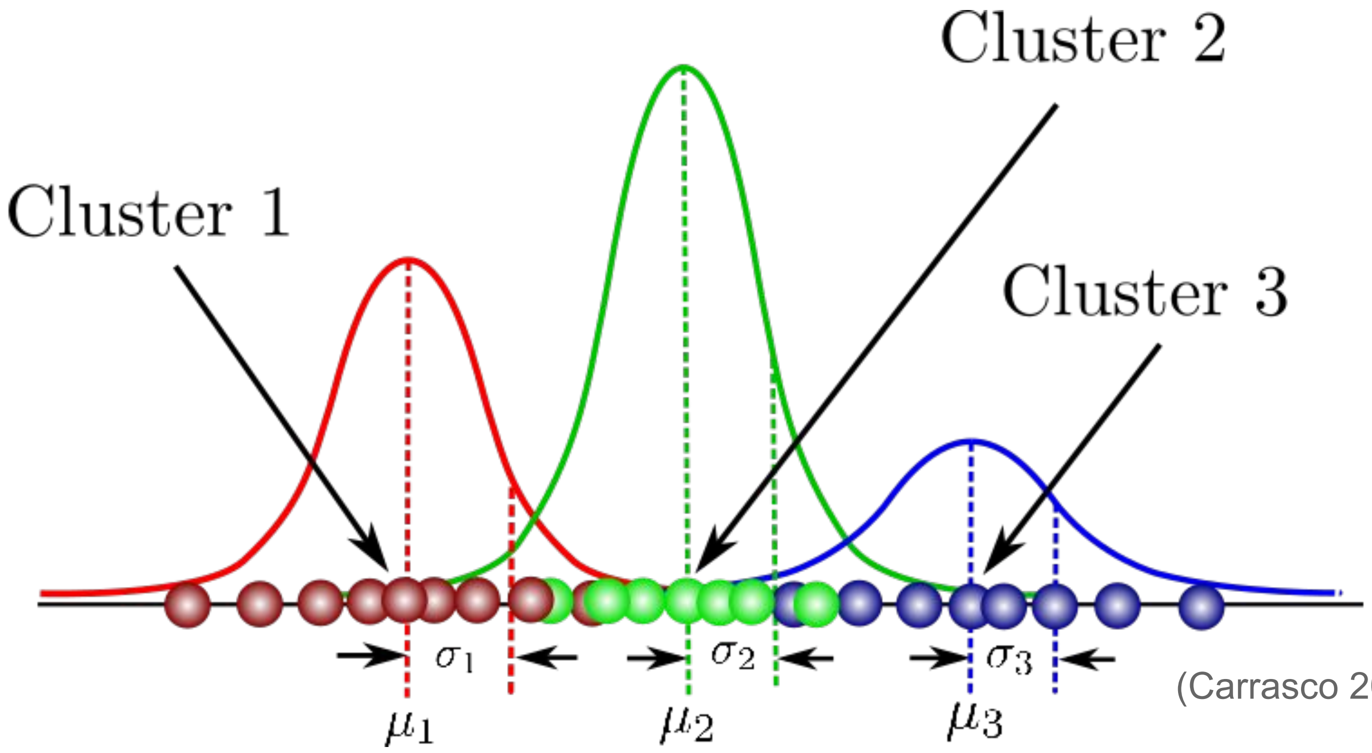
Unlabeled Clusters

Labeled Clusters



(Jeffares 2019)

Gaussian Mixture Models (GMM)



(Carrasco 2019)

Gaussian Mixture Models (cont.)

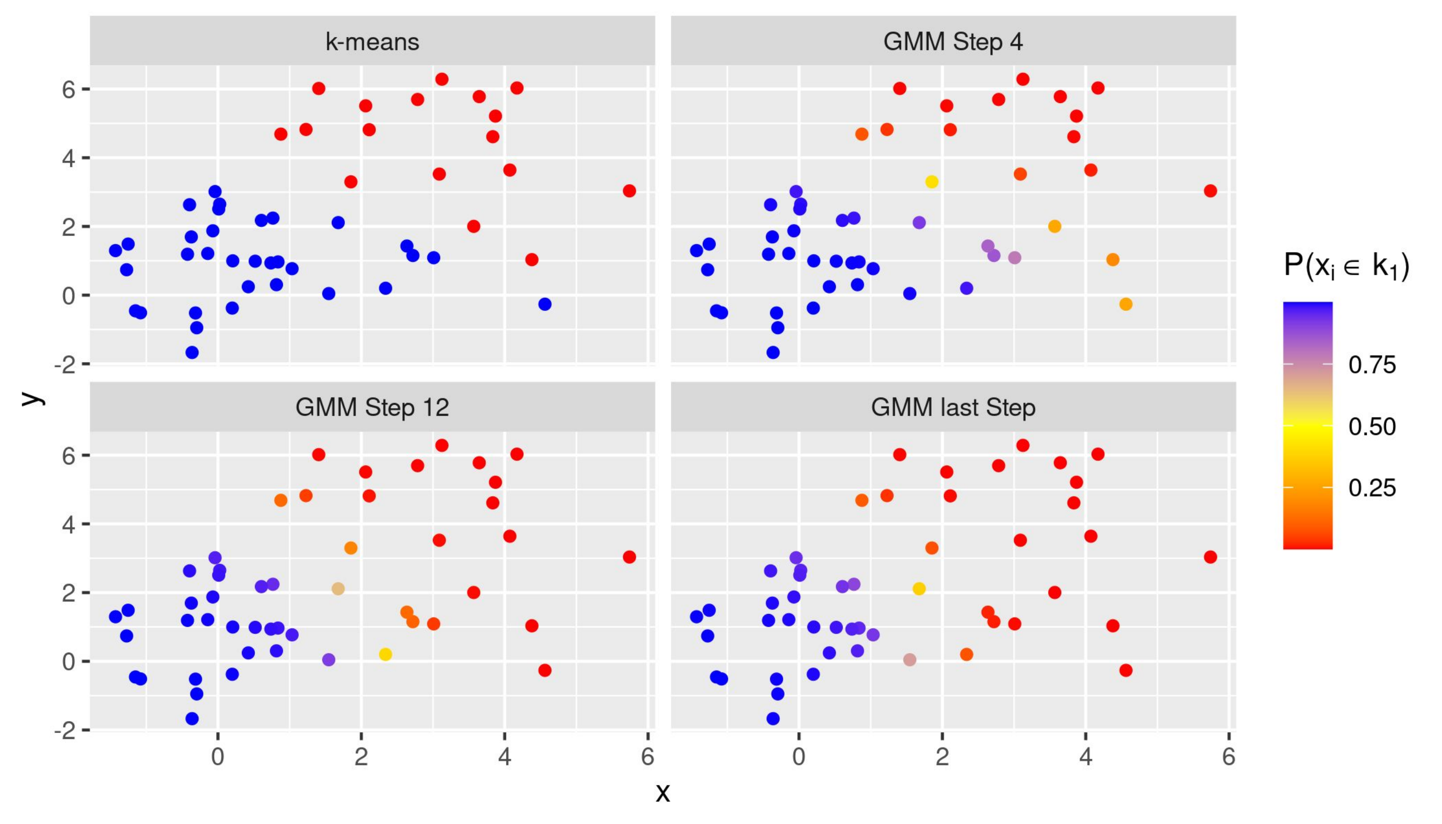
$$P(X|\mu, \Sigma, \pi) = \sum_{k=1}^{K} \pi_k N(X|\mu_k, \Sigma_k)$$

- π_k is the mixing probability for cluster k
- $N(X|\mu_k, \Sigma_k)$ is the multivariate Gaussian density for the k^{th} cluster with mean vector μ_k, and variance-covariance matrix Σ_k
- K is the number of clusters

Expectation Maximization (EM)

1. Initialize data with *k*-means
2. E-step: creates “soft labels” for each observation
3. M-step: uses soft labels from E-step to estimate parameters
4. Repeat until convergence

Expectation Maximization for Gaussian Mixture Models



Mclust{mclust}

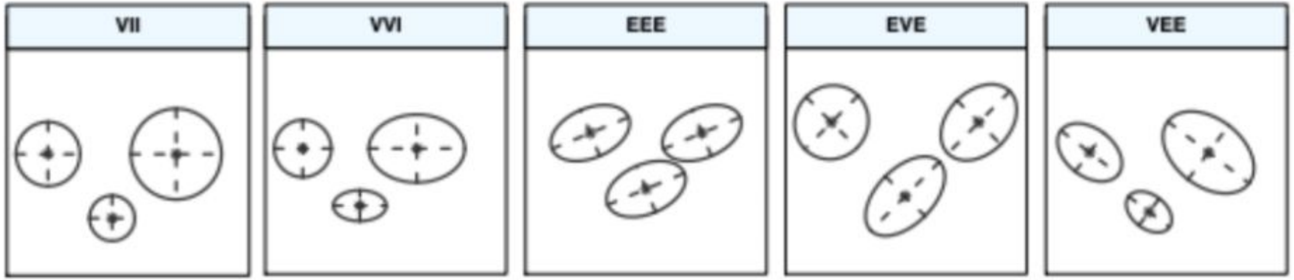
$$\Sigma_k = \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T$$

- λ_k determines the volume of each cluster
- $\mathbf{A}_k$ is a diagonal matrix that controls the shape of each cluster with the requirement that the determinant equals to one;
- $\mathbf{D}_k$ determines the orientation of each cluster, so whether the clusters are aligned with the coordinate axes, not aligned with coordinate axes, or allowed to vary in their orientation; it is an orthogonal matrix

Model examples:

(Scrucca et al. 2016)

Model	Σ_k	Distribution	Volume	Shape	Orientation
VII	$\lambda_k I$	Spherical	Variable	Equal	–
VVI	$\lambda_k \mathbf{A}_k$	Diagonal	Variable	Variable	Coordinate axes
EEE	$\lambda \mathbf{D} \mathbf{A} \mathbf{D}^T$	Ellipsoidal	Equal	Equal	Equal
EVE	$\lambda \mathbf{D} \mathbf{A}_k \mathbf{D}^T$	Ellipsoidal	Equal	Variable	Equal
VEE	$\lambda_k \mathbf{D} \mathbf{A} \mathbf{D}^T$	Ellipsoidal	Variable	Equal	Equal



k-means vs. EM for GMM

k-means

- Not model based
- Primary goal is to identify clusters
- Highly sensitive to initial centroids
- Data needs to be standardized

- Iterative process
- Results in clusters

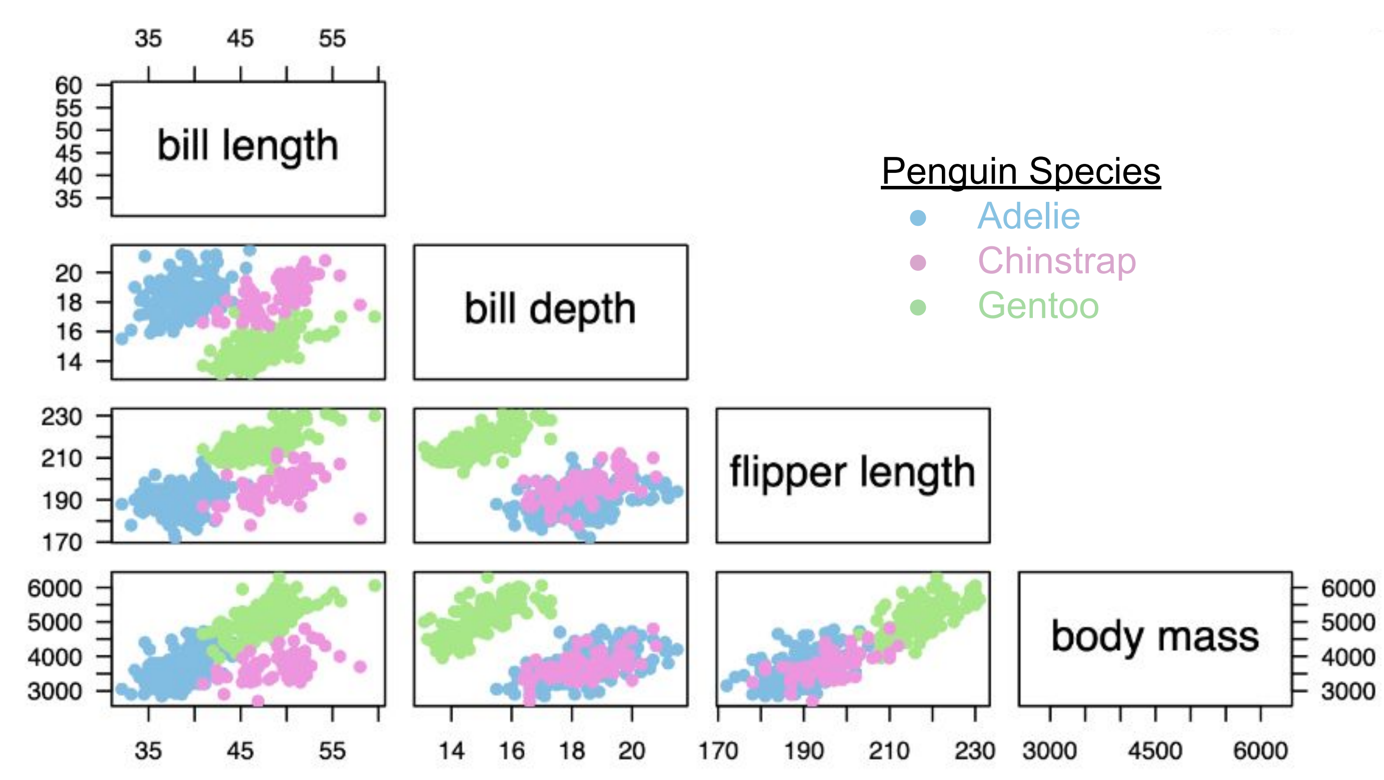
EM for GMM

- Model based
- Primary goals are to estimate parameters (mean, variance, etc.) and identify clusters

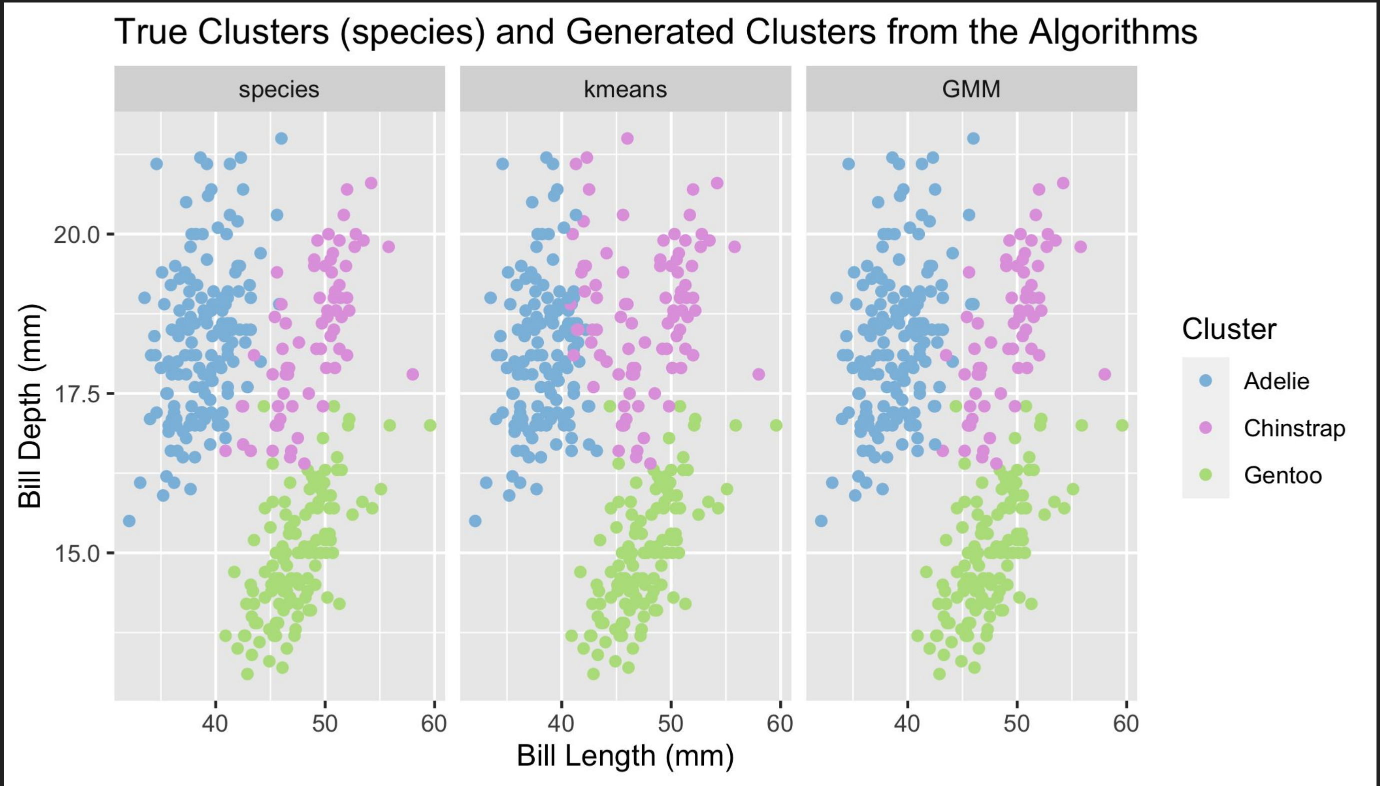
Case study: `palmerpenguins`

(Horst, Hill, and Gorman 2020)

Case study: `palmerpenguins`



Case study: `palmerpenguins`



Case study: `palmerpenguins`

	k-means			GMM		
	Cluster 1	Cluster 2	Cluster 3	Cluster 1	Cluster 2	Cluster 3
Adelie	127	24	0	151	0	0
Chinstrap	5	63	0	5	63	0
Gentoo	0	0	123	0	0	123

- Average misclassification for *k*-means: 100*(24+5)/342 = 8.48%
- Average misclassification for GMM: 100*(5)/342 = 1.46%

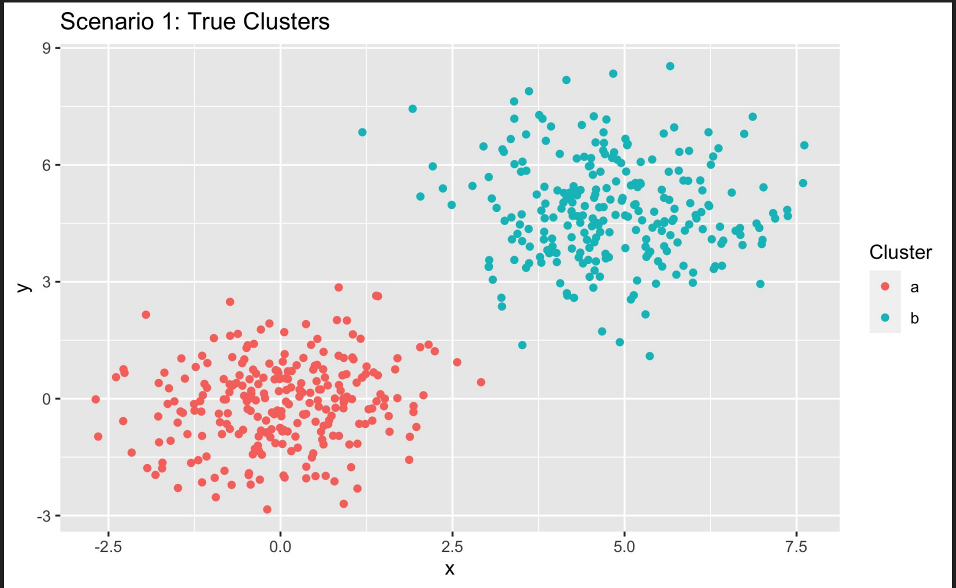
Simulation study:

- Generated data according to two multivariate Gaussian distributions
- 50 datasets per scenario

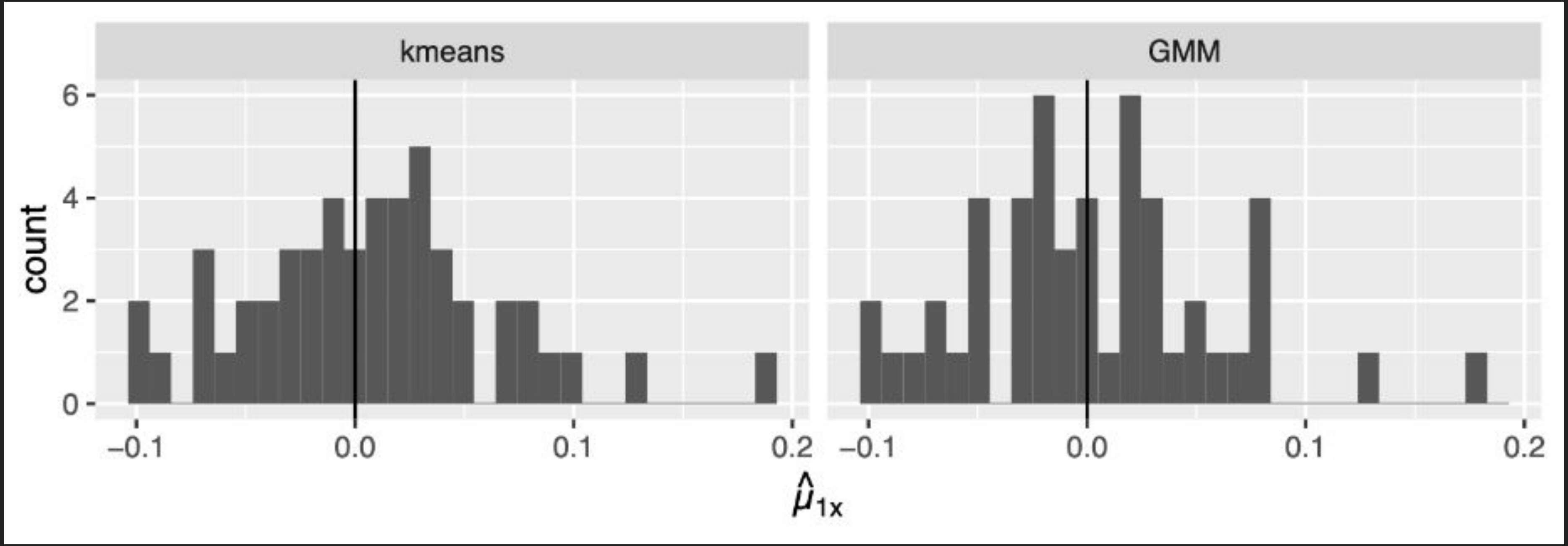
Simulation study: Scenario 1

Table 1: Scenario 1 - Model VII with $\Sigma_k = \lambda_k I$

Parameter	Cluster 1	Cluster 2
$\mu_k = \begin{bmatrix} \mu_{kx} \\ \mu_{ky} \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 5 \\ 5 \end{bmatrix}$
Σ_k	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 1.5 & 0 \\ 0 & 1.5 \end{bmatrix}$
n_k	250	250



Simulation study: Scenario 1 results



Simulation study: Scenario 1 results

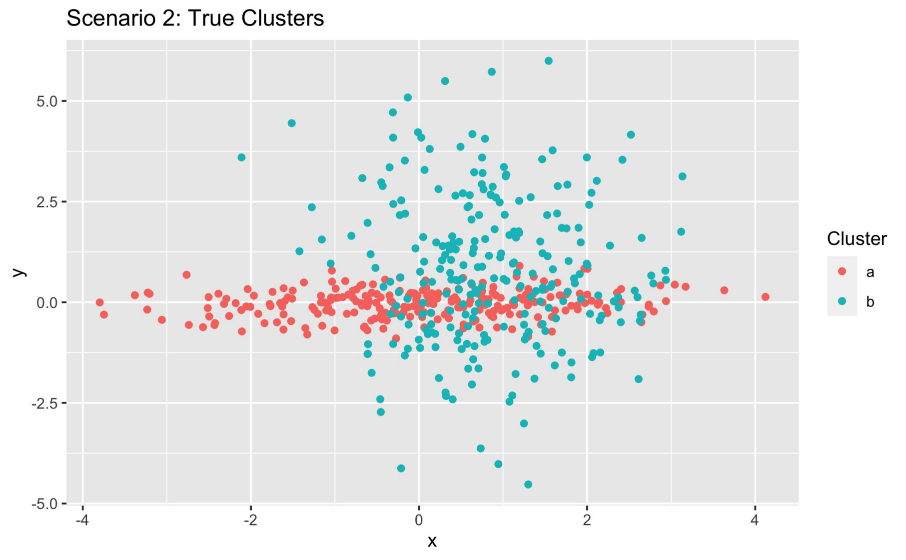
	k-means		GMM	
	Cluster 1	Cluster 2	Cluster 1	Cluster 2
Group A	249.98	0.02	249.82	0.18
Group B	0.8	249.2	0.36	249.64

- Average misclassification for *k*-means: 0.164%
- Average misclassification for GMM: 0.108%

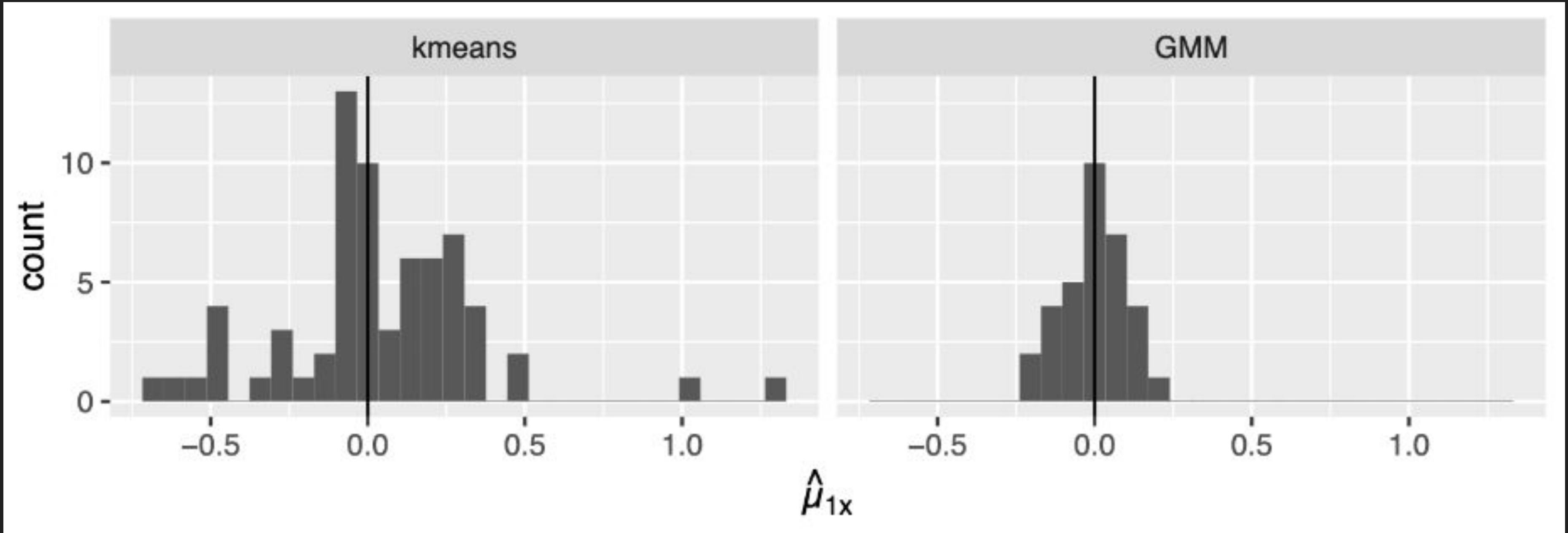
Simulation study: Scenario 2

Table 2: Scenario 2 - Model VVI with $\Sigma_k = \lambda_k A_k$

Parameter	Cluster 1	Cluster 2
$\mu_k = \begin{bmatrix} \mu_{kx} \\ \mu_{ky} \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 1 \end{bmatrix}$
Σ_k	$\begin{bmatrix} 2 & 0 \\ 0 & 0.1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix}$
n_k	250	250



Simulation study: Scenario 2 results



Simulation study: Scenario 2 results

	k-means		GMM	
	Cluster 1	Cluster 2	Cluster 1	Cluster 2
Group A	217.98	32.02	234.7	15.3
Group B	107.98	142.02	52.48	197.52

- Average misclassification for *k*-means: 28%
- Average misclassification for GMM: 13.56%

Conclusion and Future Work

- Both scenarios indicated that GMM classified data more accurately than *k*-means but as mentioned before, *k*-means is more widely used than EM for GMM
- Some limitations: only two models out of the fourteen possible models were tried and the data were simulated according to a Gaussian distribution
- Scenario in which *k*-means classified data more accurately than GMM

References

Carrasco, Oscar Contreras. 2019. "Gaussian Mixture Models Explained." *Towards Data Science.* https://towardsdatascience.com/gaussian-mixture-models-explained-6986aaf5a9.

Horst, Allison Marie, Alison Presmanes Hill, and Kristen B Gorman. 2020. *Palmerpenguins: Palmer Archipelago (Antarctica) Penguin Data.* https://allisonhorst.github.io/palmerpenguins/

Jeffares, Alan. 2019. "K-Means: A Complete Introduction." *Towards Data Science.* https://towardsdatascience.com/k-means-a-complete-introduction-1702af9cd8c.

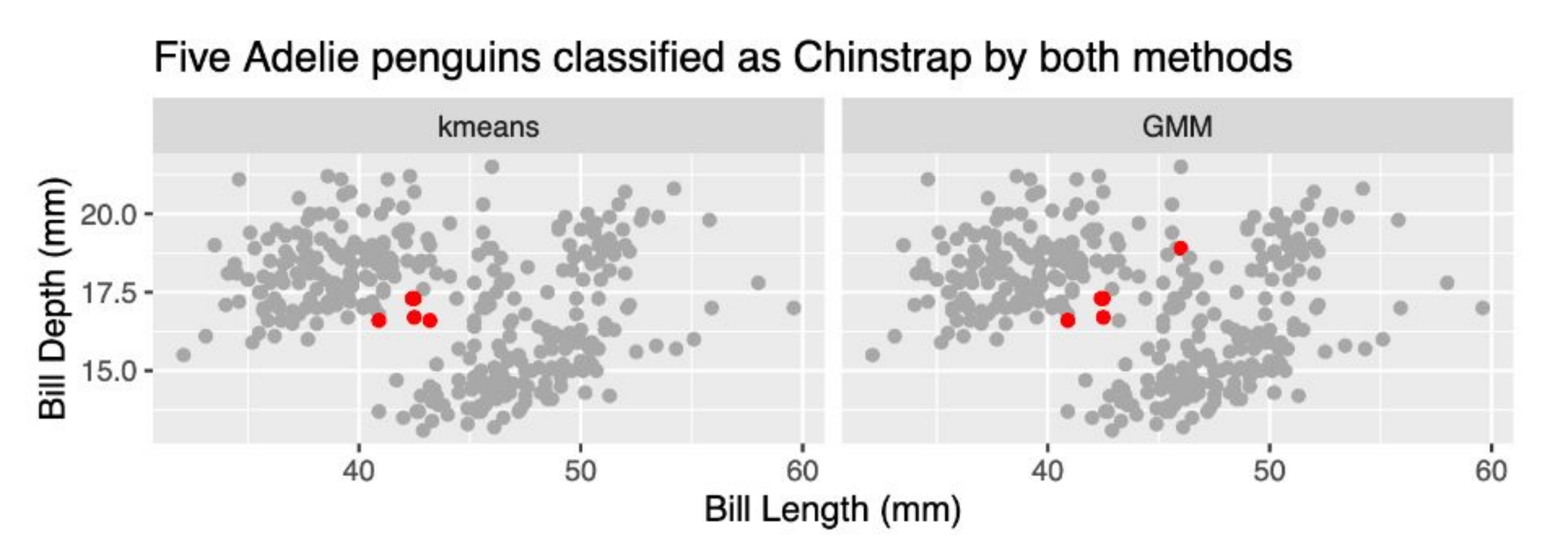
Scrucca, Luca, Michael Fop, T. Brendan Murphy, and Adrian E. Raftery. 2016. "mclust 5: Clustering,Classification and Density Estimation Using Gaussian Finite Mixture Models." *The R Journal* 8 (1): 289–317. https://doi.org/10.32614/RJ-2016-021

Questions?

All code & the paper can be found at my Github:

https://github.com/emmaleda/MS-project-kmeans-and-GMM

Case study: `palmerpenguins`



Simulation study: unusual parameter estimates

$\hat{\mu}_1$	$\hat{\mu}_2$	Confusion Matrix
$\begin{bmatrix} 1.03 \\ -0.22 \end{bmatrix}$	$\begin{bmatrix} -0.43 \\ 1.21 \end{bmatrix}$	$\begin{bmatrix} 149 & 144 \\ 101 & 106 \end{bmatrix}$
$\begin{bmatrix} 1.32 \\ -0.01 \end{bmatrix}$	$\begin{bmatrix} -0.46 \\ 1.11 \end{bmatrix}$	$\begin{bmatrix} 130 & 146 \\ 120 & 104 \end{bmatrix}$