University of Mannheim
Business Informatics and Mathematics
Fall 2023
IE 675b Machine Learning
Dr. Rainer Gemulla

Assignment 4: Latent Variable Models

December 10, 2023

Laila Albalkhi
Jonah Niklas Bjørgsvik Wiecek

lalbalkh | 1968154 | albalkhl@uwindsor.ca
jwiecek | 1966868 | jwiecek@students.uni-mannheim.com

Contents

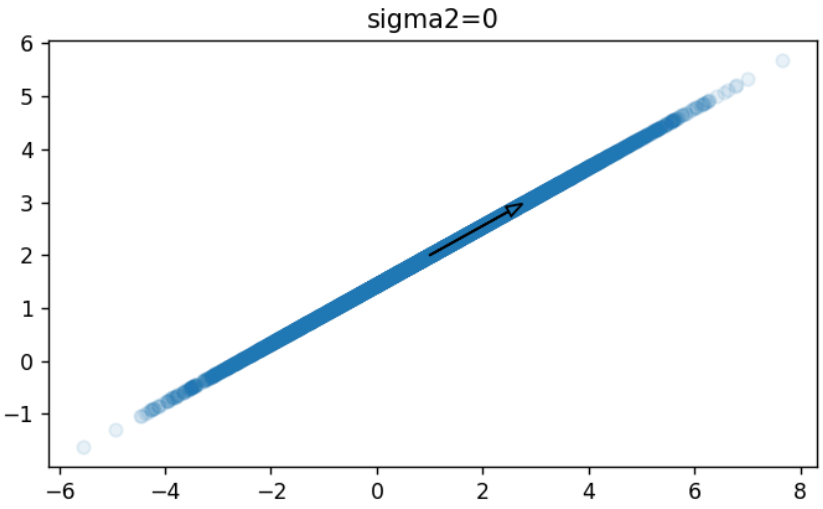


Figure 1.1: When $\sigma^2 = 0$, we have no noise in the dataset.

1 Probabilistic PCA

1.a Data Distribution

Our σ^2 term refers to the amount of noise in our data. As noise increases, the spread of the data along the vector orthogonal to the weight vector increases as well. We can interpret this as the confidence that the model has, and as noise increases, the confidence decreases. When noise is 0, we can see a straight line in the direction of the weight vector. When σ^2 is 10, on the other hand, the data is no longer as clearly aligned with the weight vector as before.

1.b Maximum Likelihood Estimation

We know that the MLE of σ^2 can be interpreted as the average variance of the discarded dimensions. In this case, we end up with an MLE of 0. We can interpret this as saying that when $L = 2$, we have 0 discarded dimensions. In other words, all of our data can be described with 2 latent variables or eigenvectors. This is clear from the fact that we are only dealing with 2 dimensions, so with 2 latent variables, we are describing the data completely.

1.c Latent Variables

Using the Scree plot, we can see a steep decline in the eigenvalues at around $L = 20$, in Figure 1.3. This correlates to the validation data we use. We have a minimum negative log-likelihood of approximately 5825 when $L = 20$. This means that 20 latent variables were used by the secret dataset.

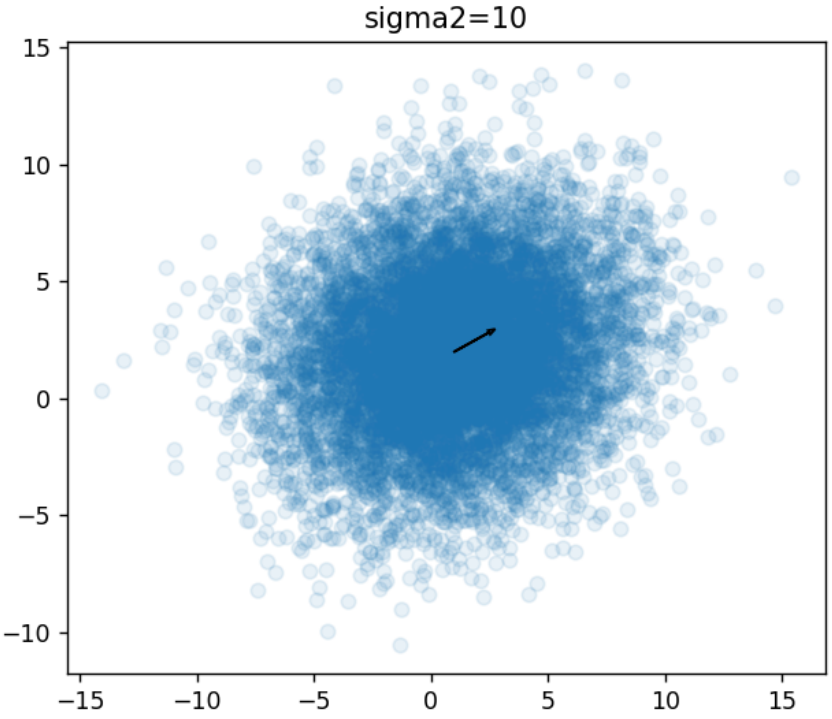


Figure 1.2: Increasing noise when $\sigma^2 = 10$

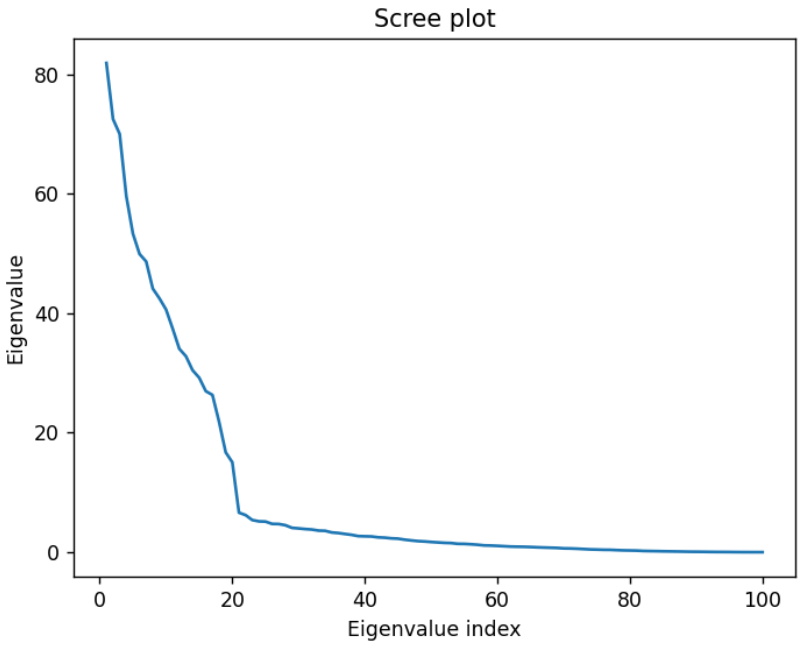


Figure 1.3: Caption

2 Gaussian Mixture Models

2.a Plot Description

We generate 5 clusters with the provided gmm_gen function. The means are:

$$[(0,0),(10,0),(-10,0),(0,10),(0,-10)] \tag{1}$$

The co-variances were provided as well. The first cluster is an isotropic Gaussian and the other four have co-variances. More interestingly, the cluster membership probabilities π are the following:

$$[0.1,0.2,0.25,0.1,0.35] \tag{2}$$

where π can also be interpreted as cluster sizes. We also see that the third and fifth clusters have some overlap which will be hard to classify correctly.

2.b K-Means Clustering

K-means clustering assumes spherical, non-overlapping clusters of similar sizes. This is because K-means with Lloyd's algorithm uses minimum Euclidean distance to cluster centers when classifying and average position of classified points when moving the centroids.

Due to these assumptions, we can expect some problems when classifying the smaller clusters correctly and the overlapping clusters. Therefore the results were expected. We see 5 centroids, one for each cluster. The pink section in the middle, which was previously described using its own Gaussian, now shares points from the neighbouring dark green and blue distributions. This is because k-means has no understanding or notion of non-overlapping clusters of different sizes, and only cares about minimizing the Euclidian distances from the centroids.

2.c Fitting GMM

This task was implementation only. Please refer to the code provided for further information.

2.d GMM vs K-Means

Since GMM does not assume spherical, non-overlapping clusters of similar size, we don't expect the problems mentioned in section 2.b. From the plot of the labelled points after running gmm_fit we see that it is very similar to the 5 Gaussian distributions the data is generated from. We have hard decision boundaries since we are picking the $argmax$ of the cluster membership probabilities, but the centroids are still placed according to each Gaussian distribution's mean, as expected.

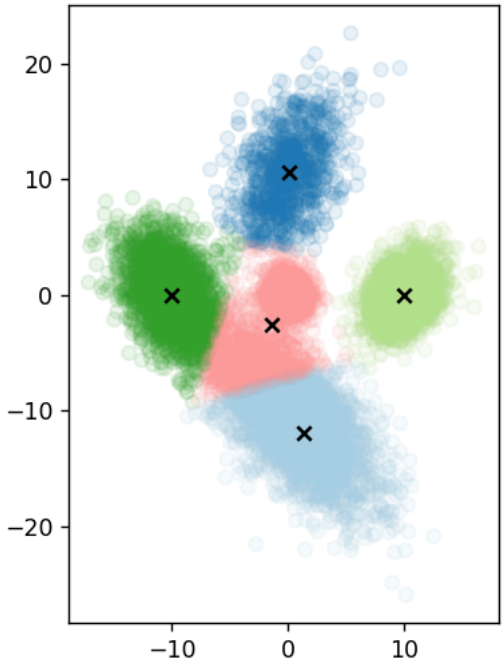


Figure 2.1: K-Means Clustering with $k = 5$

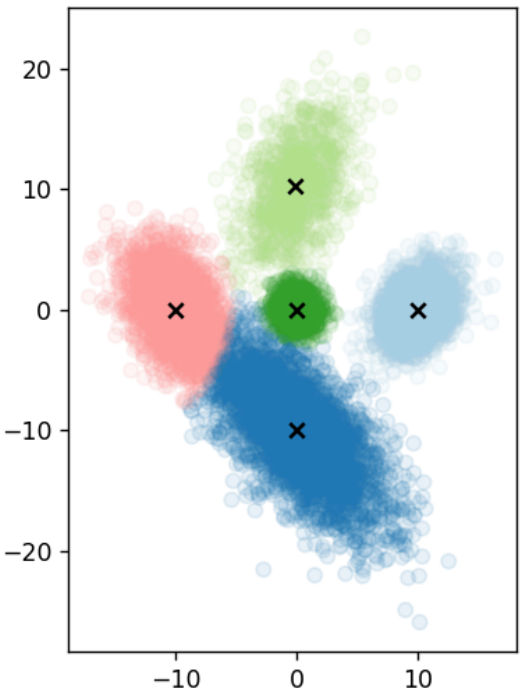
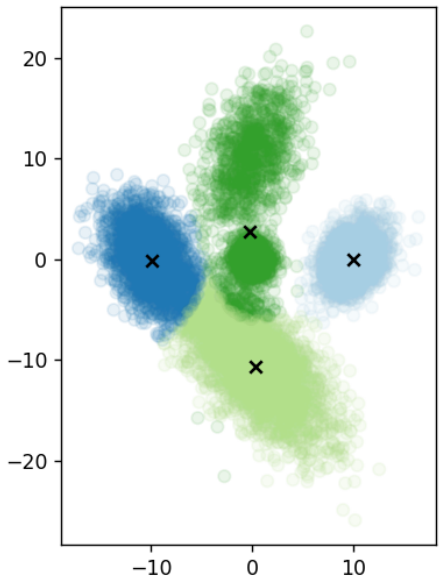


Figure 2.2: GMM with $K = 5$

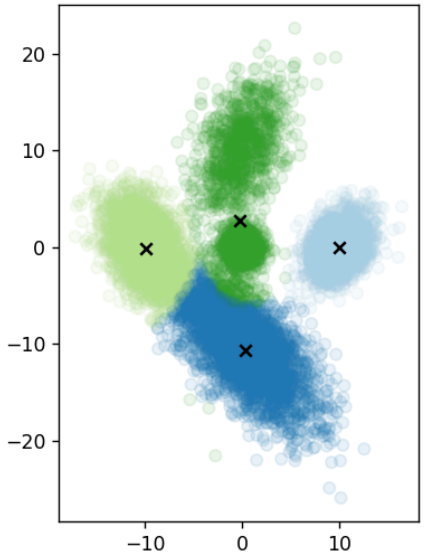
2.e K-Means with Different K Values

When $K = 4$, we are fitting fewer clusters than we have Gaussian distributions that make up the data. For this reason, the original 5 separated distributions now are combined to make 4. We see that the solutions are highly dependent on random initialization. It tends to merge two of the original clusters and try to explain the data with one Gaussian. When we run multiple random iterations, we can see this in Figure 2.3.

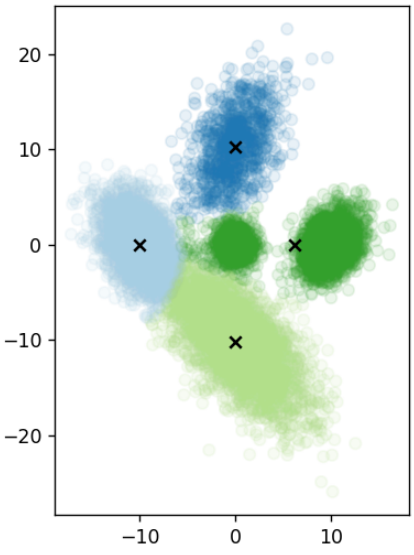
When $K = 6$, we are fitting more clusters than we have Gaussian distributions. In some of the solutions, we see that we fit two clusters with very similar means and covariances, one may have a higher scale than the other on the covariance. This leads to one assigning to the closer points and the other (with scaled-up covariance) assigning some of the outer points. What we can see is two almost entirely overlapping centroids in some of the solutions. Oftentimes, this 6th cluster does not have many data points that are assigned to it.



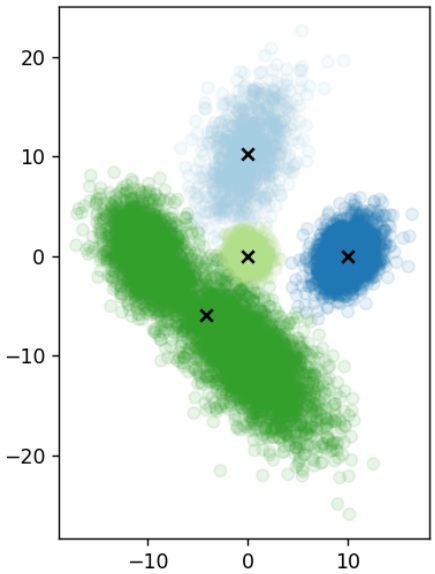
(a) Random Iteration 1



(b) Random Iteration 2

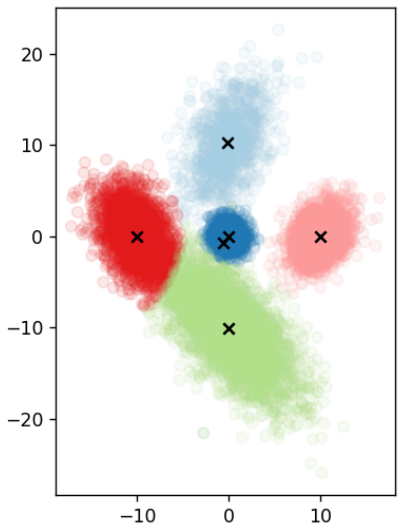


(c) Random Iteration 3

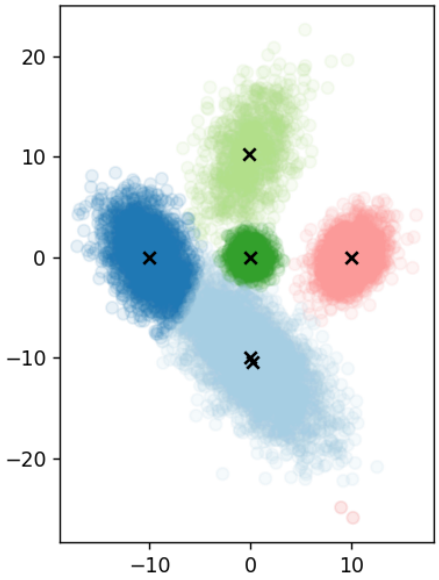


(d) Random Iteration 4

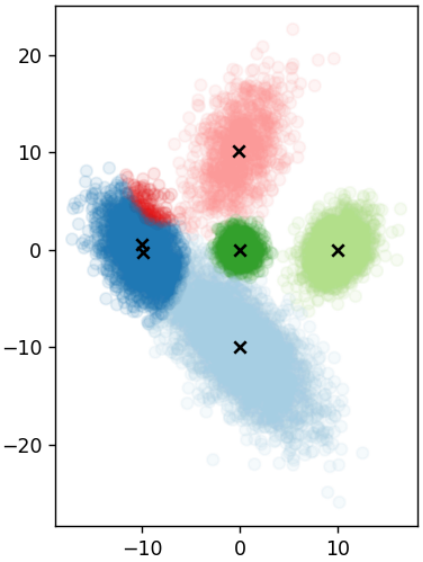
Figure 2.3: GMM with $K = 4$



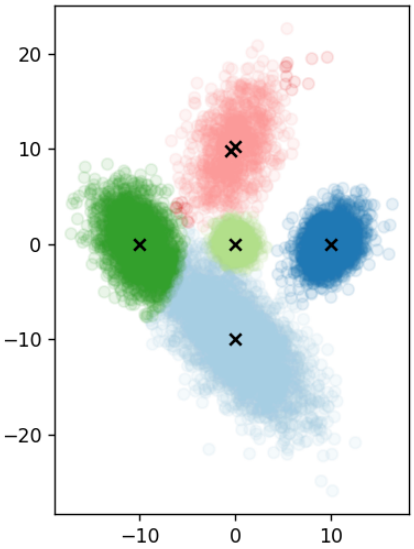
(a) Random Iteration 1



(b) Random Iteration 2



(c) Random Iteration 3



(d) Random Iteration 4

Figure 2.4: GMM with $K = 6$