ENCM 509 Lab 3

February 1, 2019

Group Members: Andrew Schneider, Patrick Settle

Instructor: Dr. Svetlana Yanushkevich

Introduction

The purpose of this lab is to get familiar with calculating, drawing and using multidimensional gaussian curves to estimate the likelihood of a sample belonging to the same class as the curve. A multidimensional gaussian curve is described by a series of mean values for each dimension, along with a $D \times D$ covariance matrix where D is the number of dimensions and the matrix is equivalent to the variance of a one dimensional gaussian curve.

In this lab gaussian curves that represent signature populations will be estimated using two seperate techniques: supervised learning via the bayesian classifier, and unsupervised learning using the expectation maximization algorithm. In the supervised case, a set of data containing pressure and velocity means for two distinct classes will be used as training data. In the unsupervised case, a set of data containing pressure and velocity means will be used, however which class each sample belongs to will not be known. Instead, the expectation maximization algorithm will use the number of classes represented in the dataset as input.

Finally, the two dimensional gaussian curves estimated using both techniques will be compared for similarity and to potentially identify any underlying mechanisms which would lead to differences in these estimates.

Exercise:

2.1 ‘A priori’ Probabilities

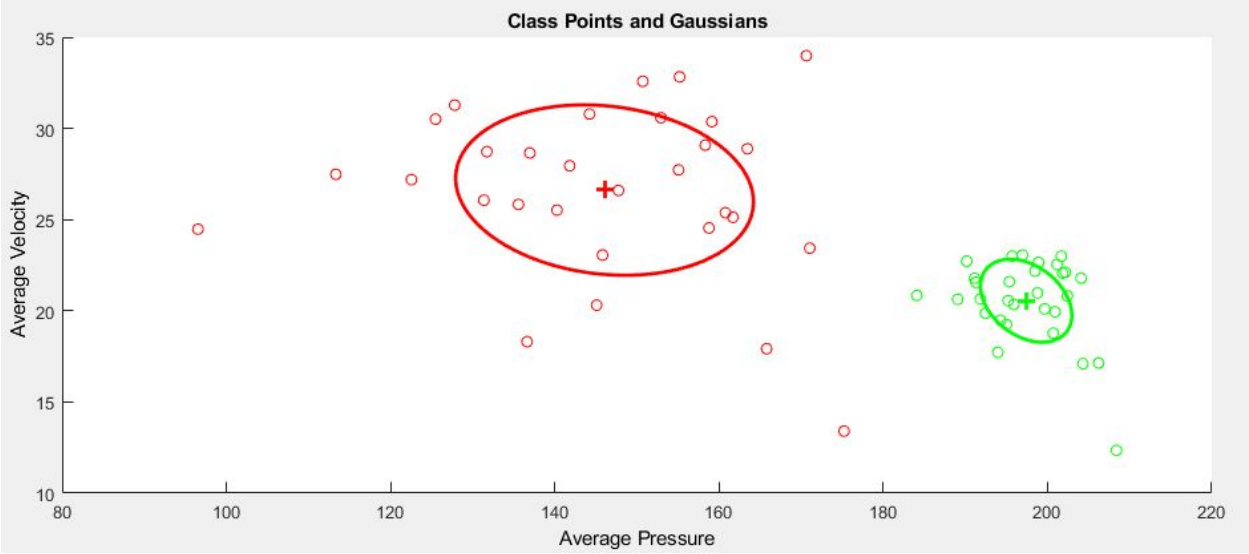

Figure 1. Plot of Each Class as Points in a 2D Plane. Authentic Signatures in Green, Forgeries in Red.

The results for μ_a, Σ_a, μ_f, and Σ_f, are summarized in the table below:

Table 1. Mean and Covariance of Authentic (a) and Forged (f) Signatures

μ_a	μ_f	Σ_a		Σ_f	
197.399038	146.054350	31.006971	-4.418137	329.015553	-11.646522
20.551676	26.618463	-4.418137	5.179986	-11.646522	21.834016

Looking at the results, there appears to be a relatively nice separation between the forged and authentic signatures. The authentic signatures appear to be written slower and with more pressure, and the forgeries are written faster with less pressure.
The means determine the centre of the 2D Gaussian functions, and the covariance determine the relative “stretching” and rotation of the Gaussians.

Looking at the covariance matrix, there appears to be much higher variance for the forged signatures. This is not very surprising, since someone writing their own signature is likely to be much more consistent than someone trying to forge it.

2.2 Bayesian Classification

The Matlab function `gloglike(point, mu, sigma)` is used to compute the likelihoods for each class.
For each class we also need to compute the prior probability as follows:
$P_c = \frac{N_c}{N}$,
Where P_c is the prior probability for class 'c', N_c is the number of samples for class c, and N is the total number of samples.
So, since there are 30 authentic signatures and 30 forged signatures, we have:
$P_a = \frac{N_a}{N}$
$P_a = \frac{30}{60} = 0.5$
$P_f = \frac{N_f}{N}$
$P_f = \frac{30}{60} = 0.5$

Table 2 shows the `gloglike` values for the authentic and forged sets.

Table 2. Gloglike Values for Authentic and Forged Signatures

Authentic gloglike	Forged gloglike
-3948.831481	-2535.344221
-3845.802564	-2197.294269
-3756.11462	-2382.141104
-3542.635357	-2029.222823
-3642.350057	-2367.09957
-3599.445055	-1205.022338
-3668.05651	-1964.453225
-3508.624723	-1555.556065
-3745.919959	-836.4522871
-3610.060997	-2059.759533
-3634.796669	-1975.64616

-3502.752299	-1950.105211
-3421.592022	-1847.954725
-3366.456263	-2117.820258
-3168.937577	-1630.204455
-3175.299288	-1851.715816
-3351.196814	-1340.657641
-3657.306552	-1531.522408
-3489.187826	-1306.147344
-3198.806997	-1575.459691
-3225.078616	-1425.545986
-3082.142673	-1332.226
-3304.316417	-1011.624224
-3128.336641	-1213.644224
-3290.392344	-1782.515322
-2878.258975	-586.3630444
-3335.873643	-1466.304192
-3344.652234	-1124.240032
-3375.101136	-1073.309171
-3589.41515	-1716.520648

Figures 2 and 3 respectively contain the 2D and 3D PDFs for the set of authentic signatures, and Figures 4 and 5 respectively contain the 2D and 3D PDFs for the forged signatures.

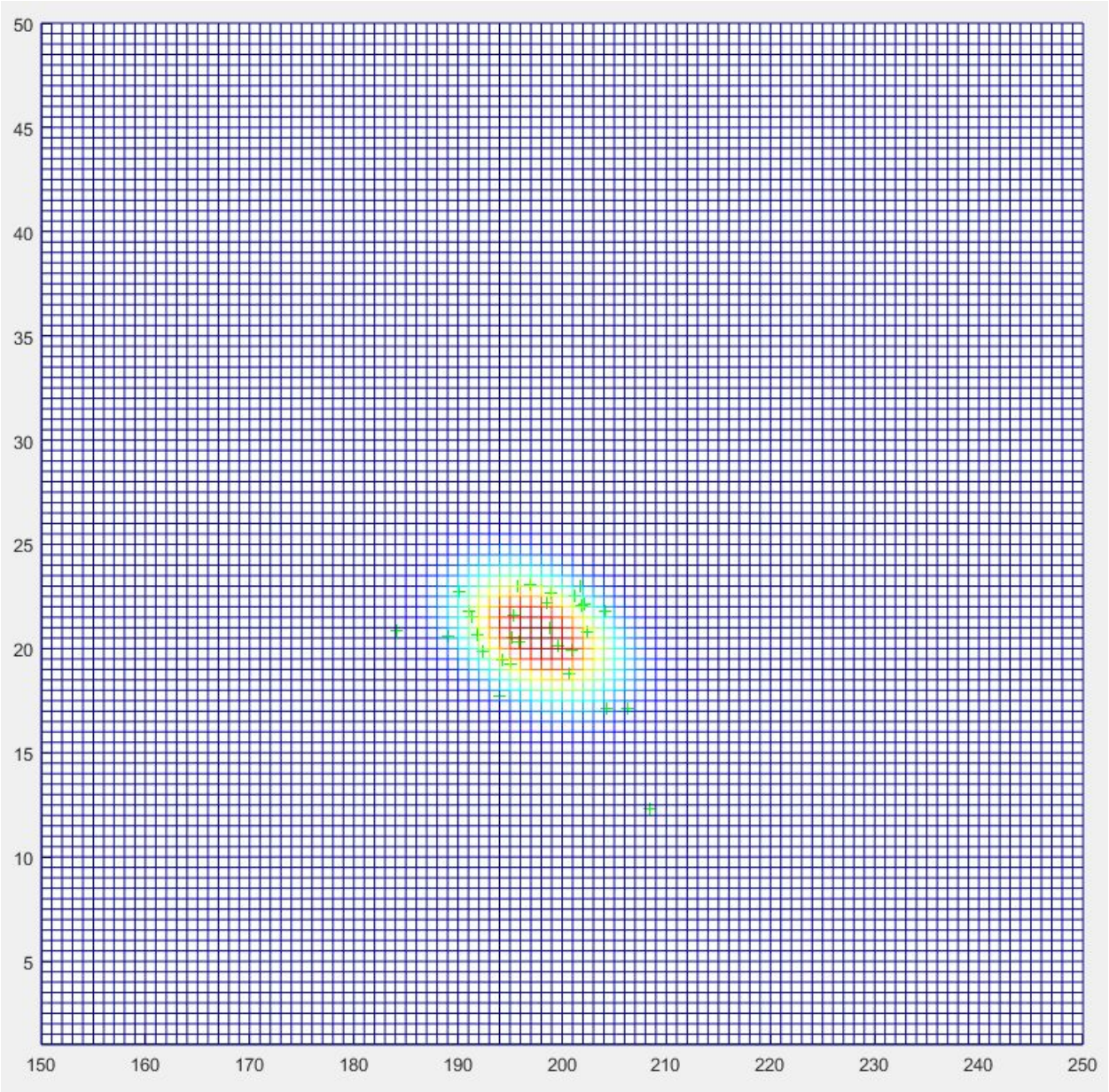

Figure 2. Two-Dimensional PDF for Authentic Signatures

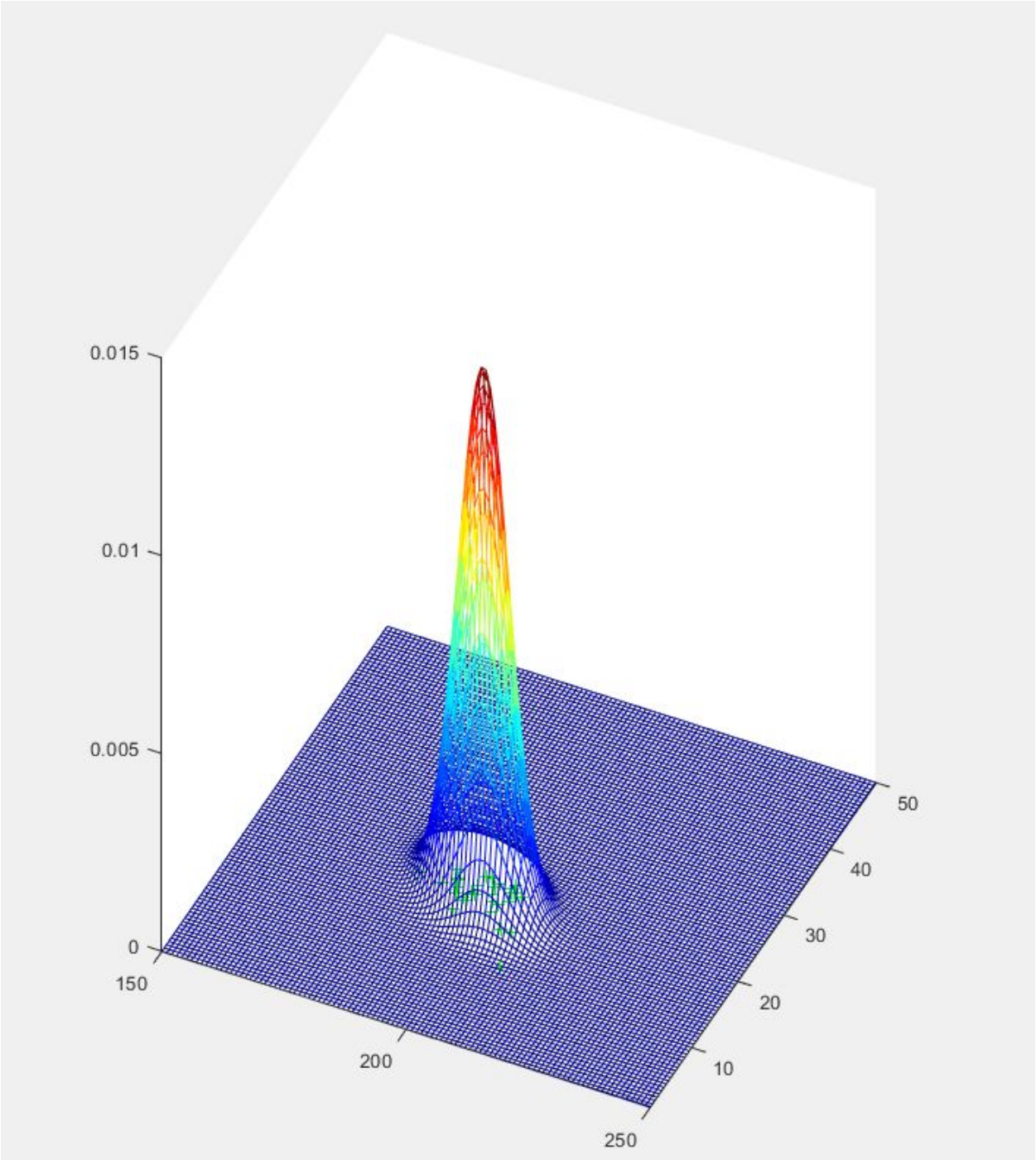

Figure 3. Three-Dimensional PDF for Authentic Signatures

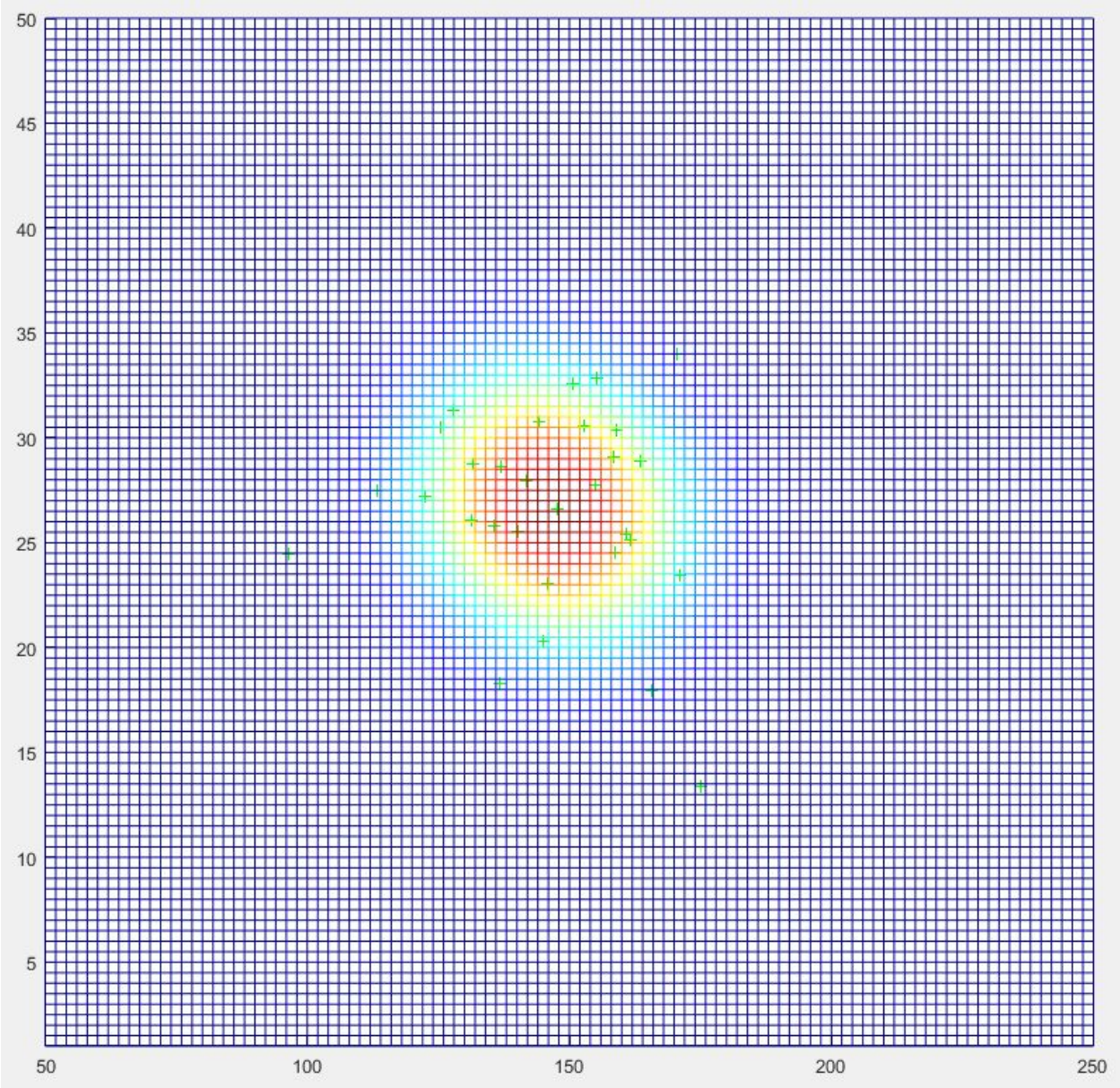

Figure 4. Two-Dimensional PDF for Forged Signatures

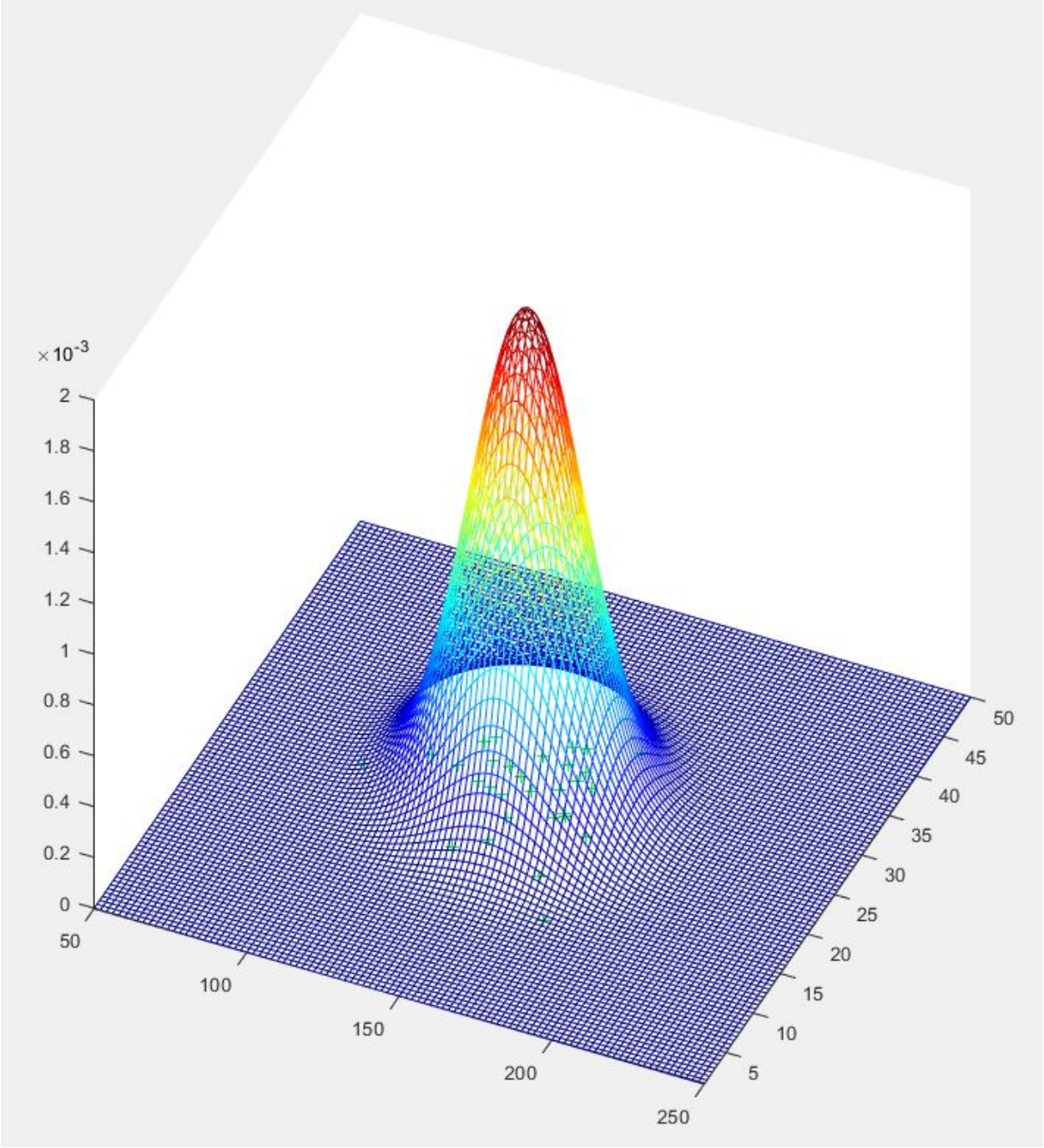

Figure 5. Three-Dimensional PDF for Forged Signatures

For the authentic signature class, the log-likelihoods computed by `gloglike` have an average of about -3446, with a standard deviation of 246.3. The forgeries have an average of about -1633, with a standard deviation of 479.9. It's not surprising that the forgery scores have a higher standard deviation, since there appeared to be much more variability in those samples.

3.2 Implementation of Expectation Maximization Algorithm

Conveniently, we can use the provided `emalgo(data, clusters)` function to perform the Expectation Maximization (EM) algorithm.

First, we can do several runs with our authentic and forged datasets using two clusters determined according to the default heuristic, where just the data and number of clusters are passed in. We'll iterate the algorithm until the total likelihood reaches an asymptotic convergence (in other words, appears to stop changing). Figure 6 shows the evolution of the clusters and the total likelihood curve as the algorithm iterates. The default colours were changed from yellow to blue for readability.

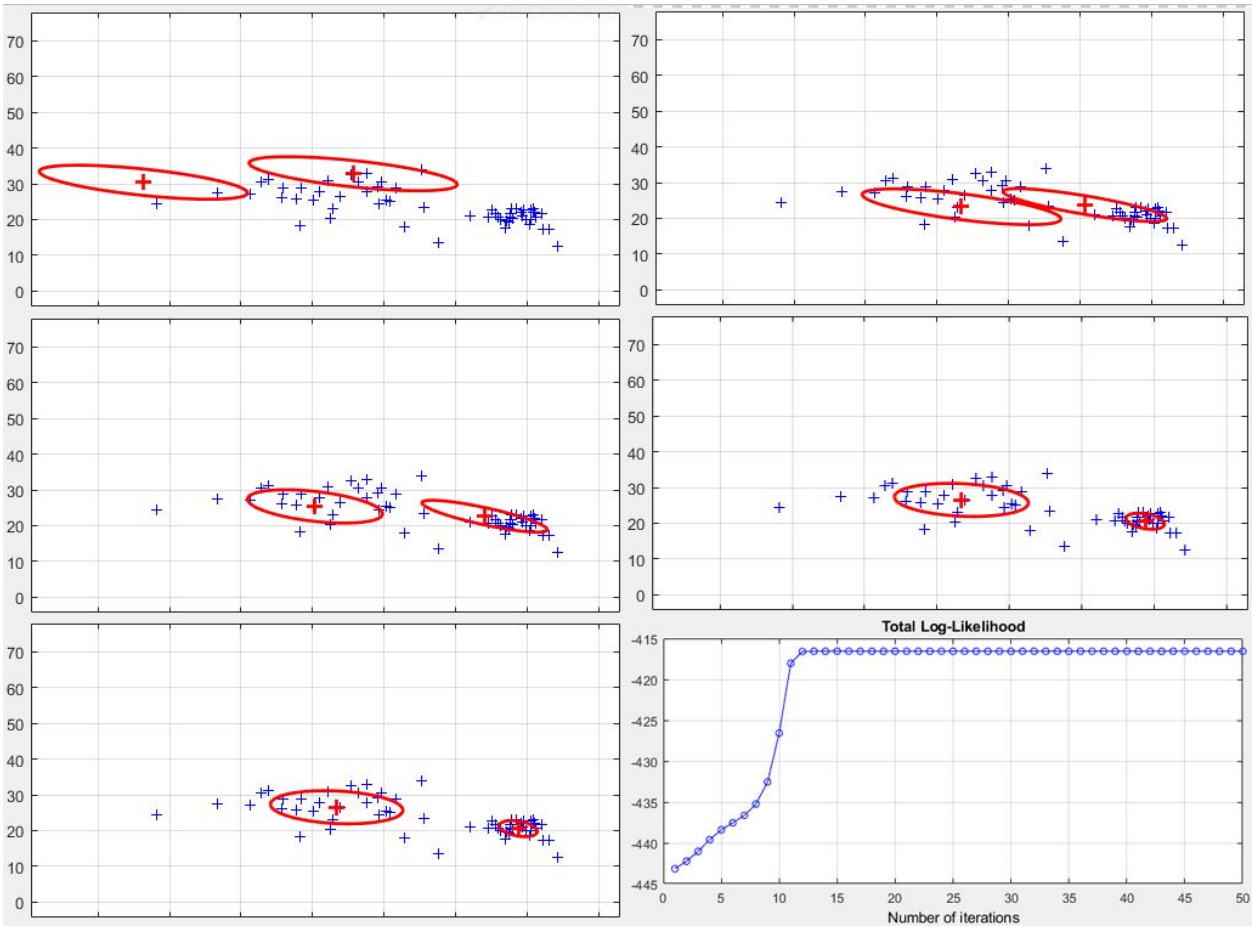

Figure 6. Evolution of Clusters and Total Likelihood Curve as EM Algorithm Iterates. Reading left ⇒ right ⇒ down from the top left, clusters after 0, 1, 6, 25, and 50 iterations. In the bottom right, the total log-likelihood curve.

The mean, variance, and prior values after the convergence of the EM algorithm are presented in Table 3 along with the values found using supervised training.

Table 3. Mean, Variance, and Prior Values - EM vs Supervised Training

Algorithm	μ_a	μ_f	Σ_a	Σ_f	$Prior_a$	$Prior_f$
Supervised Learning - Bayesian classification	197.4 20.6	146.1 26.6	31.0 -4.4 -4.4 5.2	329.0 -11.6 -11.6 21.8	0.5000	0.5000
Unsupervised Learning - Expectation Maximization	197.5 20.55	146.8 26.5	28.4 -4.2 -4.2 5.0	343.1 -15.4 -15.4 21.4	0.4918	0.5082

Brief discussion comparing the values found using either method

The mean values, covariance values and prior probabilities are all very similar between the supervised and unsupervised learning techniques. This difference may be attributed to the outlier forged sample that is closer* to the authentic populations than to its own forged population, since the unsupervised algorithm cannot be certain that the sample belongs to the forged population.

Notably the covariance matrices are more significantly different between the supervised and unsupervised learning than any other metric .

*Closer in euclidean distance to the mean of the population.

Conclusion

We estimated a 2-D gaussian probability density function using both the supervised bayesian classifier technique and the unsupervised expectation maximization algorithm. Both techniques separated our training populations into two distinct groups. The two techniques generated very similar gaussian curves, but did not generate exactly identical curves.

Appendices

The code used for this lab is included along with the submission in the folder AndyPatrickLab3Code.