Project 4 - LDA

Blair Gemmer

CSCI 548 - Pattern Recognition

Spring 2013

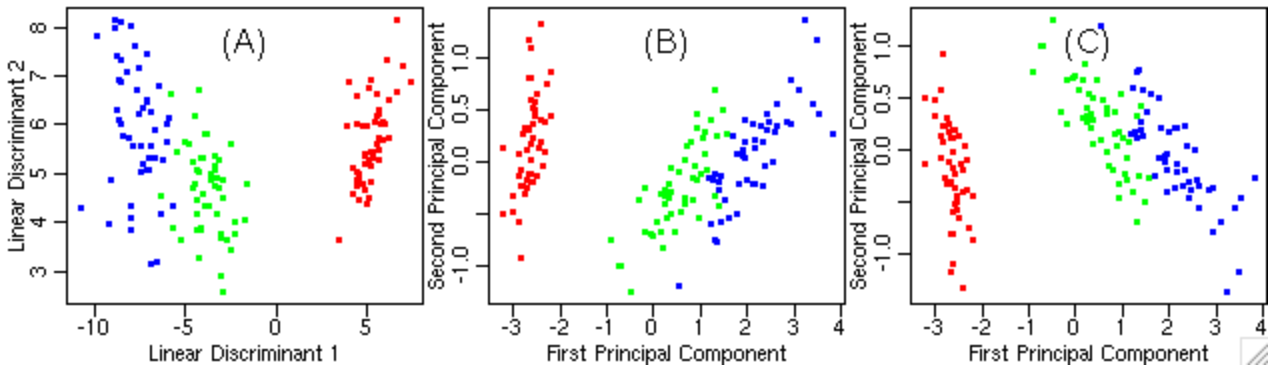

Figure 1. Iris Data (A) using Linear Discriminant Analysis, (B) using Multidimensional Scaling, and (C) using Principal Components Analysis. (KEY: Setosa = Red, Versicolor = Green, Virginica = Blue) **96% accuracy**

Answer: I believe all the figures are about even as far as separability.

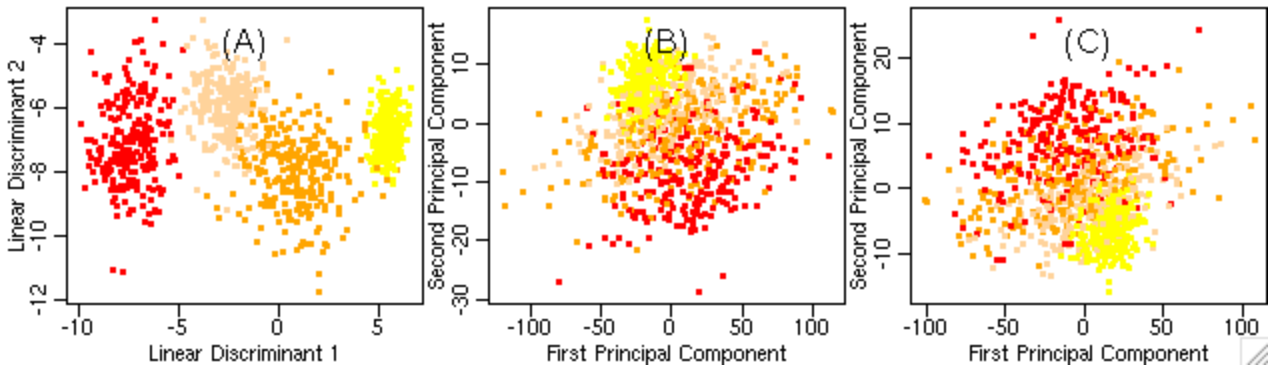

Figure 2. Fruit Data (A) using Linear Discriminant Analysis, (B) using Multidimensional Scaling, and (C) using Principal Components Analysis. (KEY: Apple = Red, Orange = Orange, Lemon = Yellow, Peach = Burlywood1) **95% accuracy**

Answer: It seems like Figure 2 (A) provides the best separability.

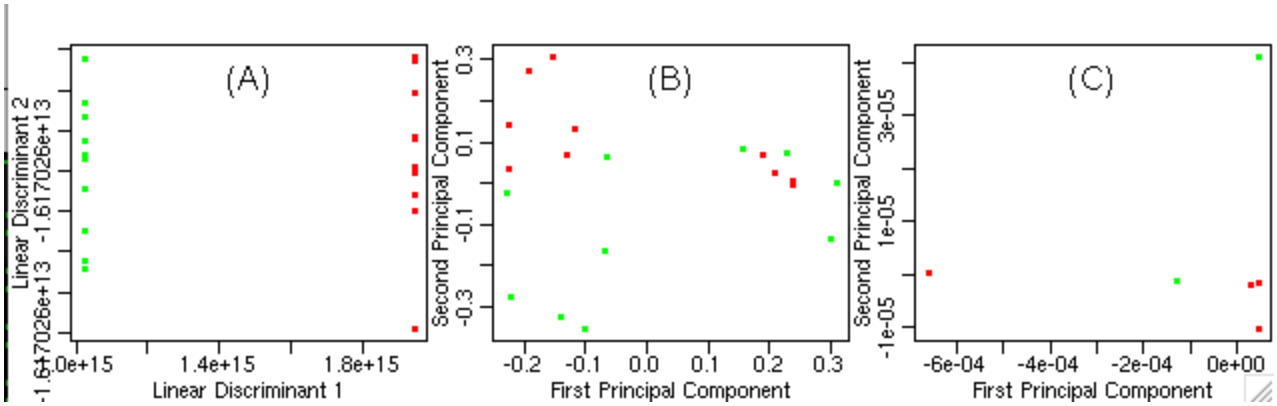

Figure 3. Mouse Data (A) using Linear Discriminant Analysis, (B) using Multidimensional Scaling, and (C) using Principal Components Analysis. (KEY: Proximal = Red, Distal = Green) **NaN% Accuracy (because test data = training data)**

Answer: I believe Figure 3, (A) provides the best separability.

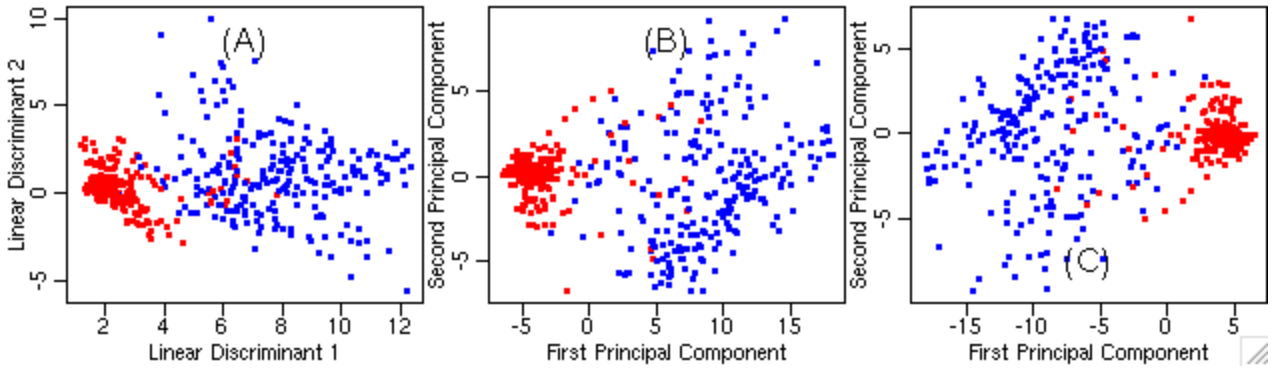

Figure 4. Cancer Data (A) using Linear Discriminant Analysis, (B) using Multidimensional Scaling, and (C) using Principal Components Analysis. (KEY: Benign = Red, Malignant = Blue) **95% accuracy**

Answer: I believe figure 4, (A) provides the best separability, but they all seem very even. It is difficult to tell.

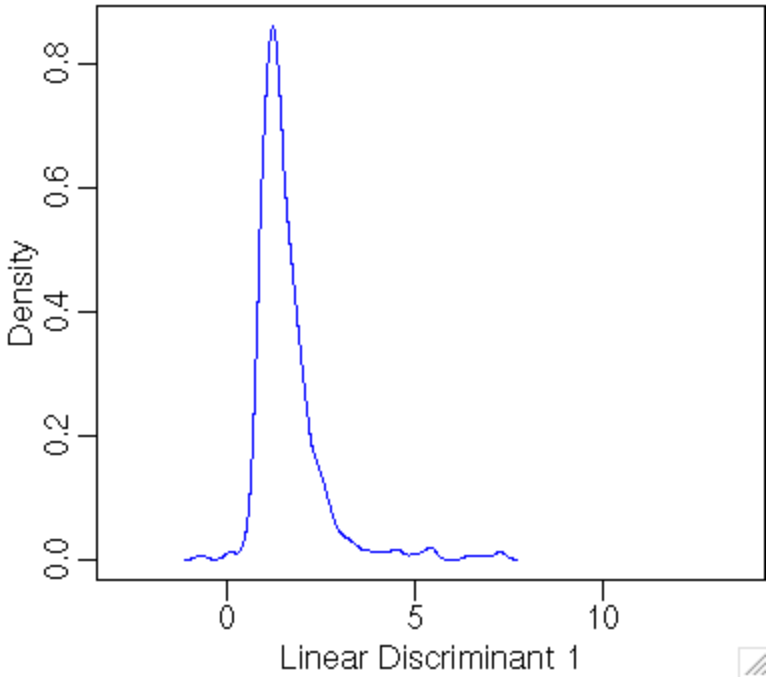

Figure 5: Density plot of the Benign tumor data.

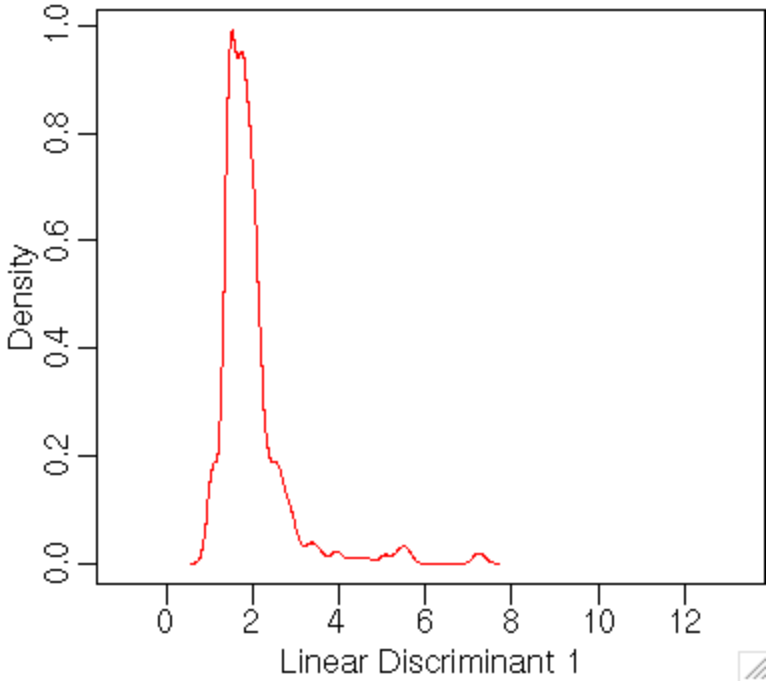

Figure 6: Density plot of the Malignant tumor data.

End Notes:

I think the LDA finds more separability because it uses a random sampling and more precise mathematical formulas to determine the separation between the types of classes. I'm not sure why some of the plots are more separated than others, but it could simply be due to the randomness of the sampling. I got much different accuracies depending on when I ran the script, because of that randomness.

As far as proximal and distal goes in the mouse data, it seems like LDA finds a huge separation in the data, but it is hard to tell since we have so few training data and we use the same training data as the test data. It would be nice to get a larger training set and see what we can find.