DWD Application on Simulated Two-class Data

Xiaosun Lu

June 24, 2011

- Generate data with two classes from random normal distribution N(0, 1).
- The sample size of each group is 100, and the dimension is 1000.
- The class mean in the first dimension is shifted to +2.53 (-2.53 resp.).
- Hanwen's DWD result is almost identical to Jason's, but much faster.
- The OOB error rate of Random Forest is 0.045 , and the running time is 19.744 s.
- SVM and LDA are also performed and compared with Hanwen's DWD result.

View the original data.

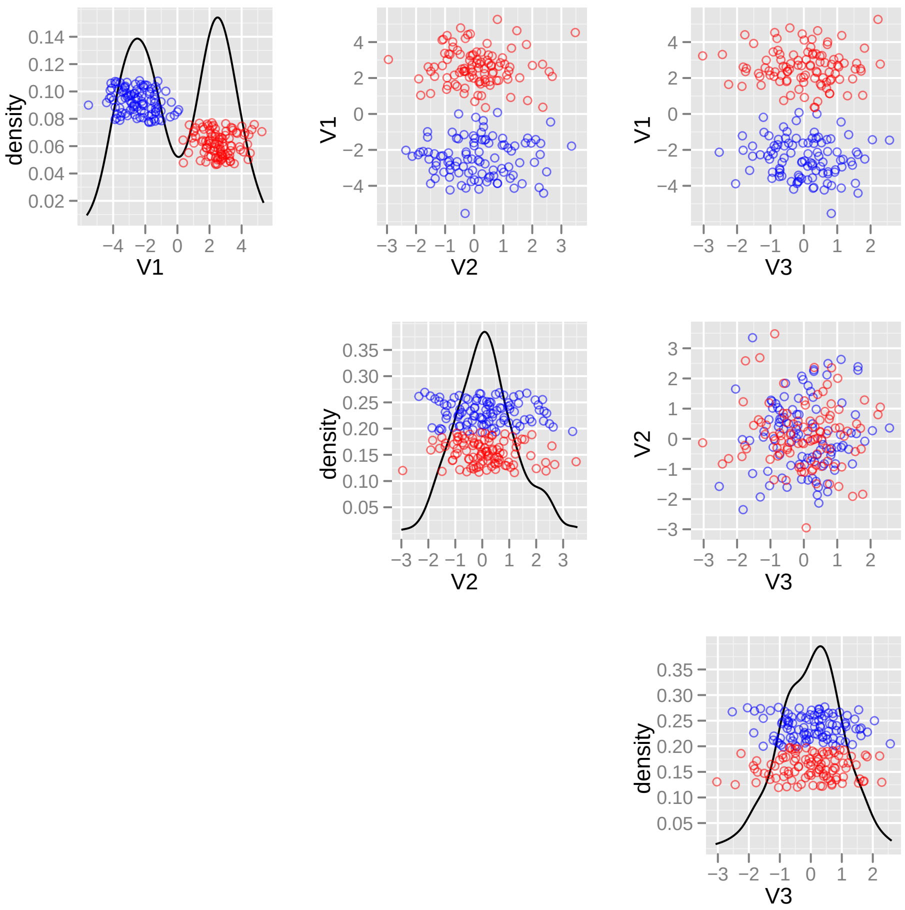

Figure 1: Scatterplots.

Rotate the point cloud onto PC directions.

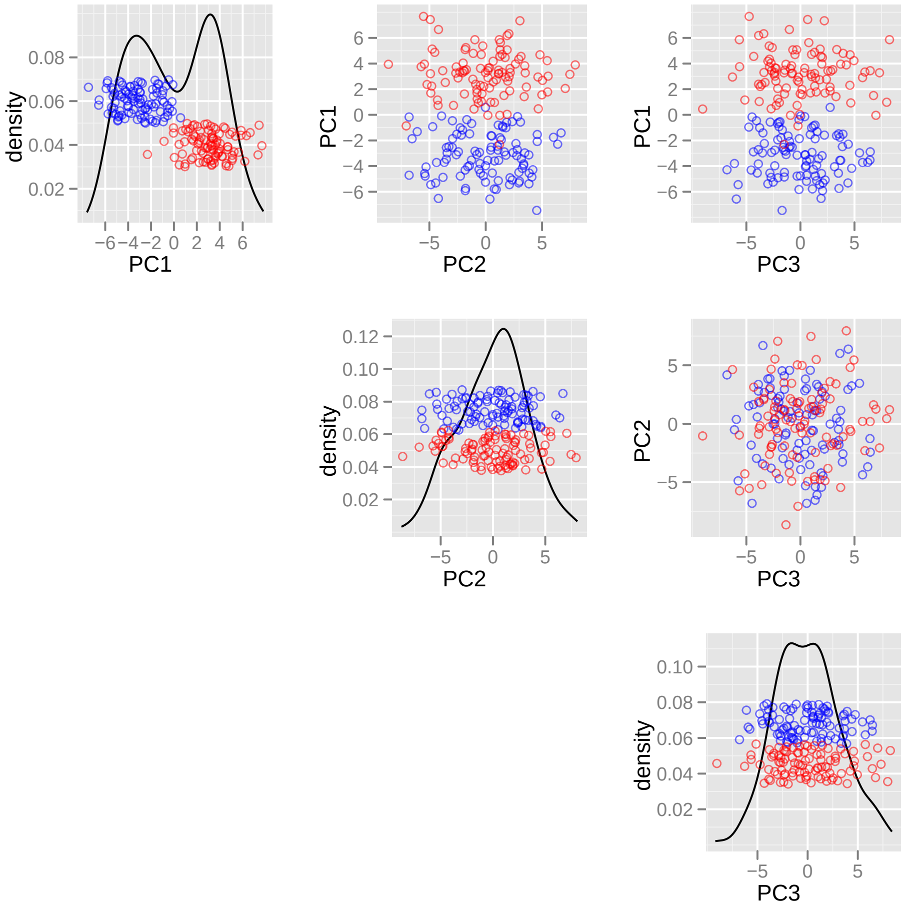

Figure 2: Data projection onto PC directions.

- Two-class DWD classification
- The angle between the DWD direction by Jason and that by Hanwen is 0.00001693 degree.
- Jason's DWD running time is 34.079 s, and Hanwen's is 6.748 s.

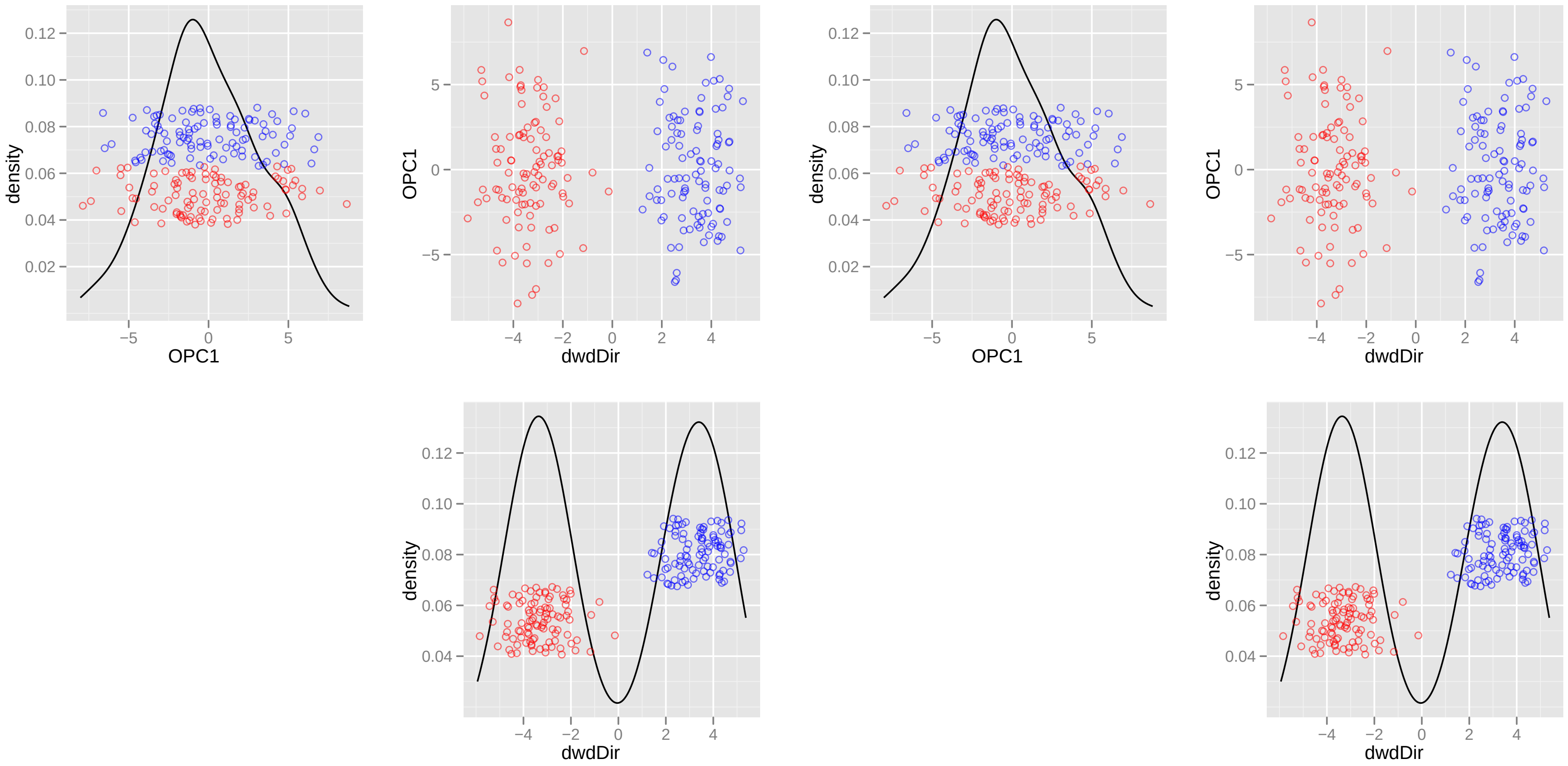

Figure 3: DWD classification. The left is Jason's result. The right is Hanwen's result.

- Cross-validation to compare Hanwen's and Jason's DWD (nrep=100, 80% trained)
- Jason's prediction error rate is 0.04 , and the 95% CI is (0, 0.09) .
- Hanwen's prediction error rate is 0.04 , and the 95% CI is (0, 0.09) .
- The angle between Jason's and Hanwen's DWD direction is 0.00005323 degree, and the 95% CI is (0.00000629, 0.00017009)

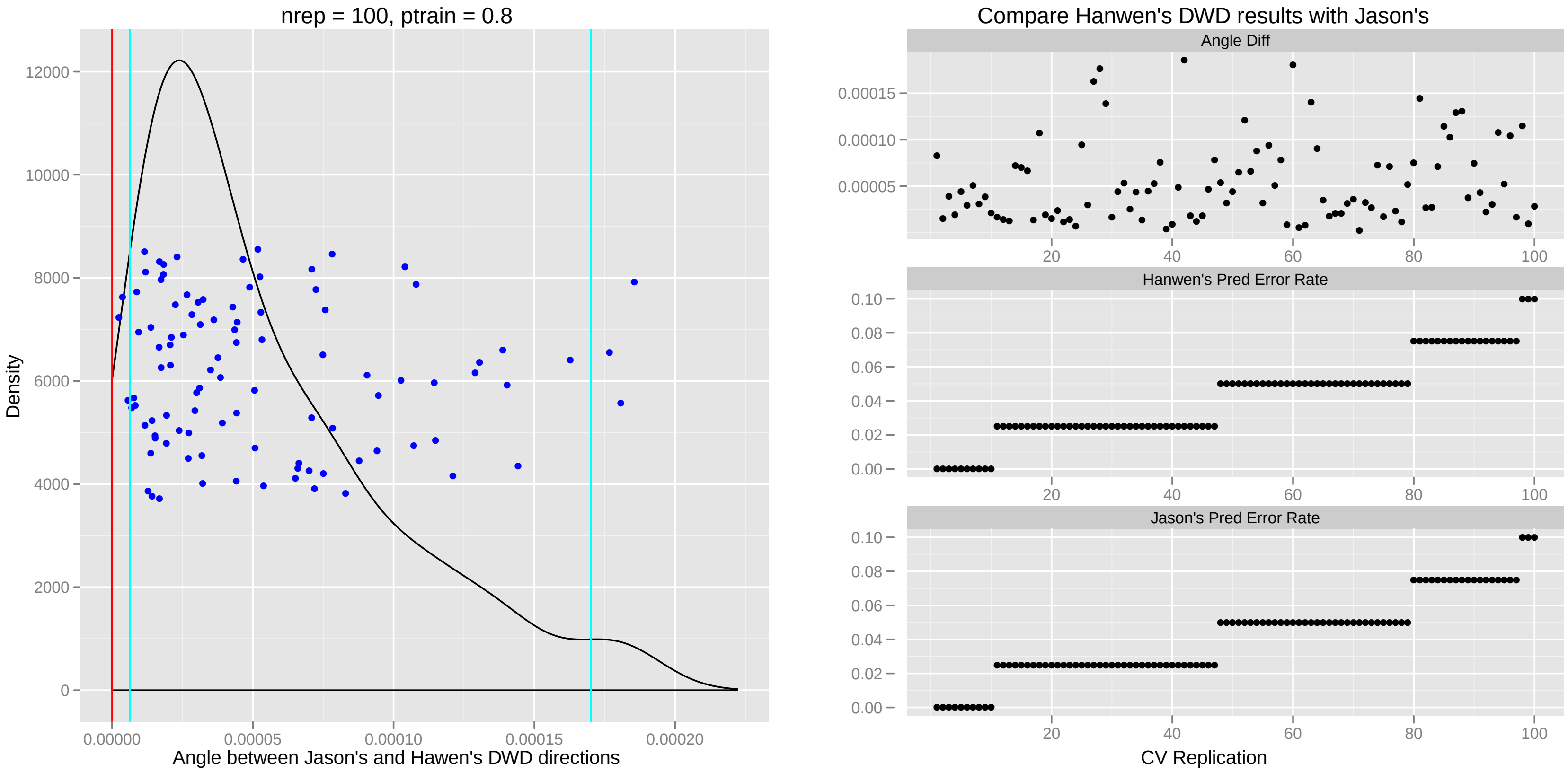

Figure 4: DWD classification. The left is the distribution of angle difference between Jason's and Hanwen's DWD directions. The right shows the angle differences and the prediction errors during the cross-validation.

- Two-class linear SVM classification (Data Piling)
- SVM running time is 0.67 s, and Hanwen's DWD running time is 6.748 s.
- CV: nrep=100, 80% trained.
- The angle between linear SVM direction and Hanwen's DWD direction is 21.89 degree, and the 95% CI is (20.33, 23.06) degree.
- The linear SVM prediction error rate is 0.07 , and the 95% CI is (0.02, 0.15)
- Hanwen's DWD prediction error rate is 0.04 , and the 95% CI is (0, 0.09) .

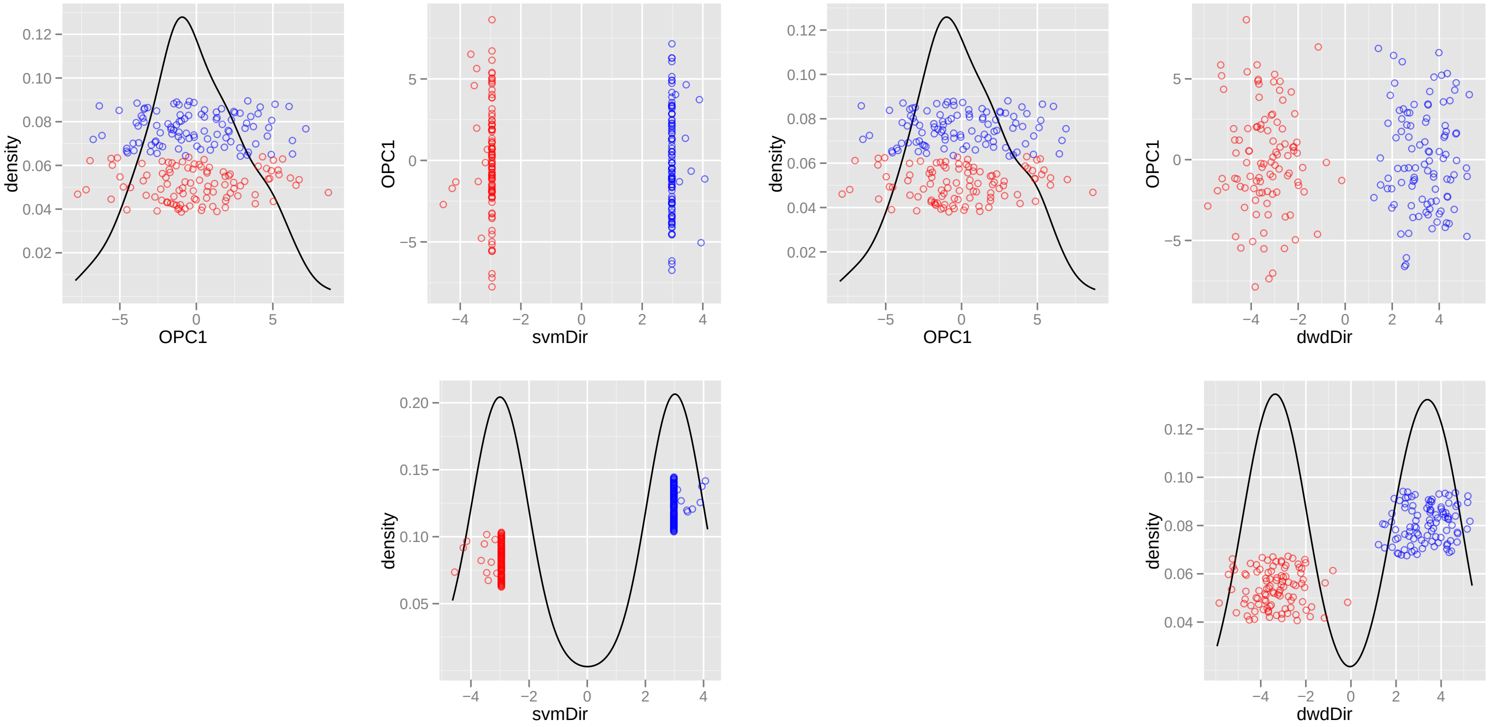

Figure 5: The left is linear SVM classification. The right is Hanwen's DWD classification.

- Two-class LDA classification
- LDA running time is 2.167 s, and Hanwen's DWD running time is 6.748 s.
- CV: nrep=100, 80% trained.
- The angle between LDA direction and Hanwen's DWD direction is 69.74 degree, and the 95% CI is (68.6, 71.07) degree.
- The LDA prediction error rate is 0.27 , and the 95% CI is (0.15, 0.41)
- Hanwen's DWD prediction error rate is 0.04 , and the 95% CI is (0, 0.09) .

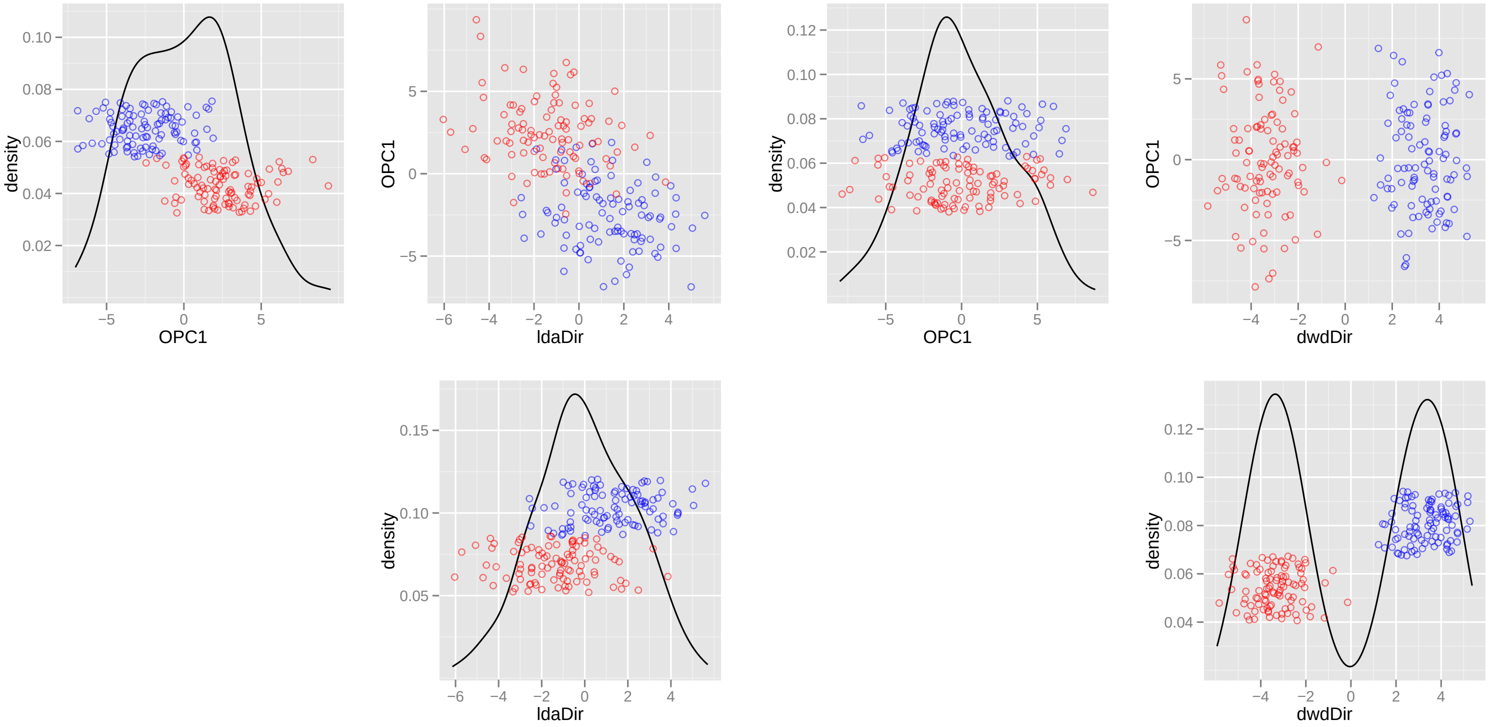

Figure 6: The left is LDA classification. The right is Hanwen's DWD classification.

- An example of prediction
- 80% training data + 20% test data
- The DWD prediction error rate is 0.05 .

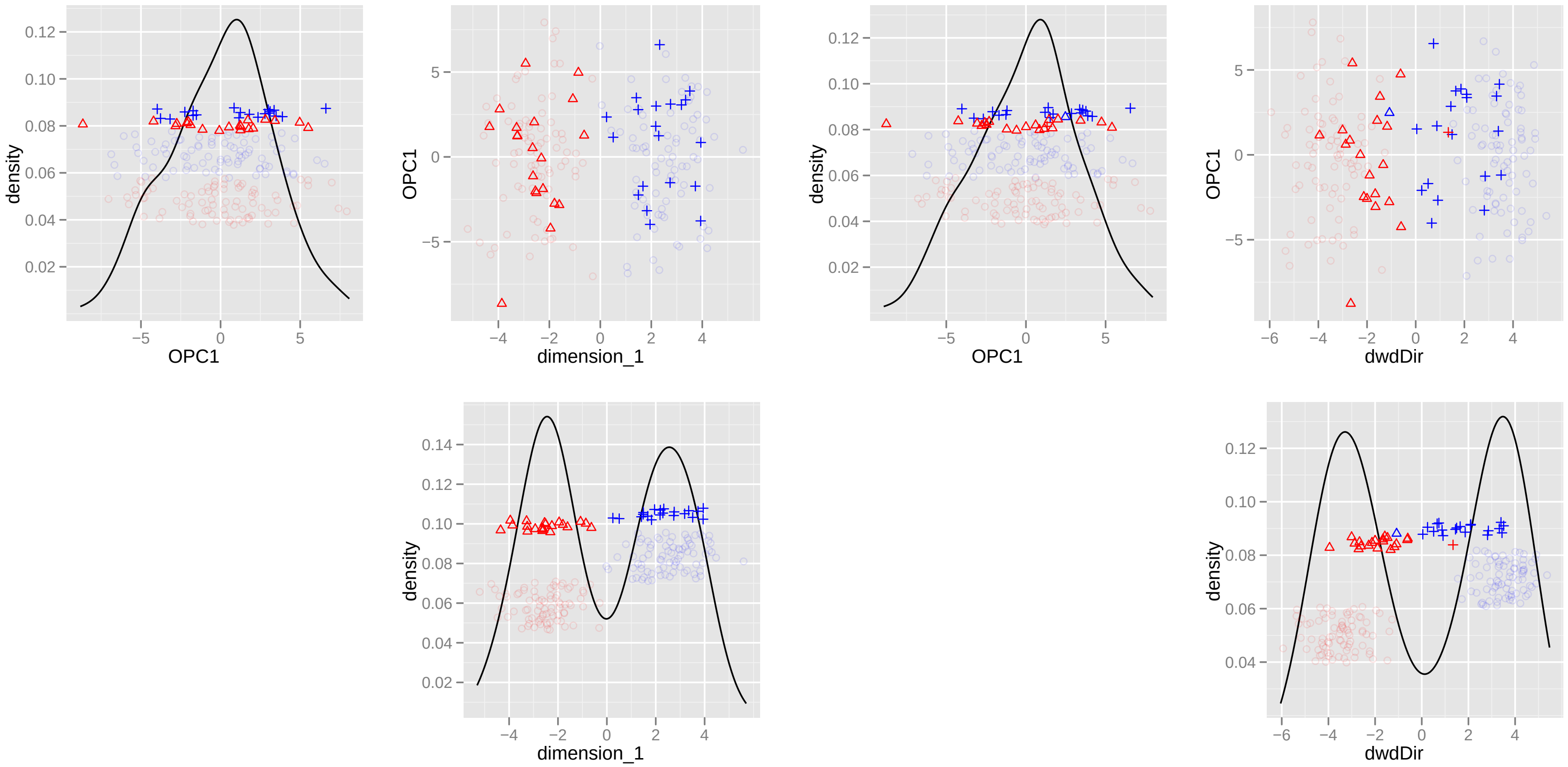

Figure 7: The left shows data on the 1st dimension. The right shows Hanwen's DWD prediction. The color indicates the true classification. The lighter colored points (circular) are from the training data, and the rest from the test data. The symbol of the points from the test data indicates the predicted class. The corss symbol corresponds to the blue group, and the triangle correspondsto the red group.

- The same example of prediction: 80% training data + 20% test data
- DWD is better than SVM in prediction.
- The SVM prediction error rate is 0.1 , and the DWD error rate is 0.05 .

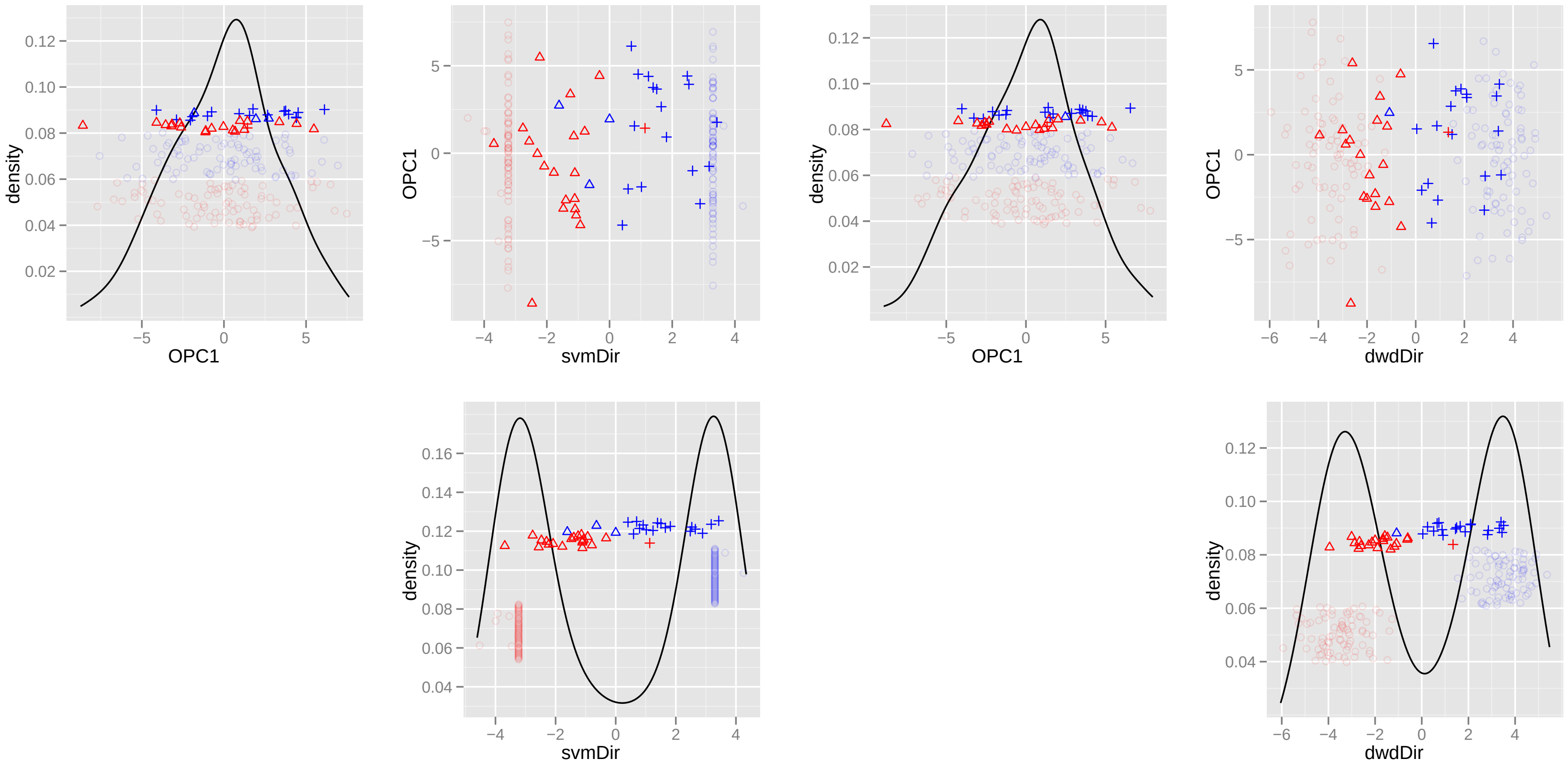

Figure 8: The left shows linear SVM prediction. The right shows Hanwen's DWD prediction. The color indicates the true classification. The lighter colored points (circular) are from the training data, and the rest from the test data. The symbol of the points from the test data indicates the predicted class. The corss symbol corresponds to the blue group, and the triangle correspondsto the red group.

- The same example of prediction: 80% training data + 20% test data
- DWD is better than LDA in prediction.
- The LDA prediction error rate is 0.175 , and the DWD error rate is 0.05 .

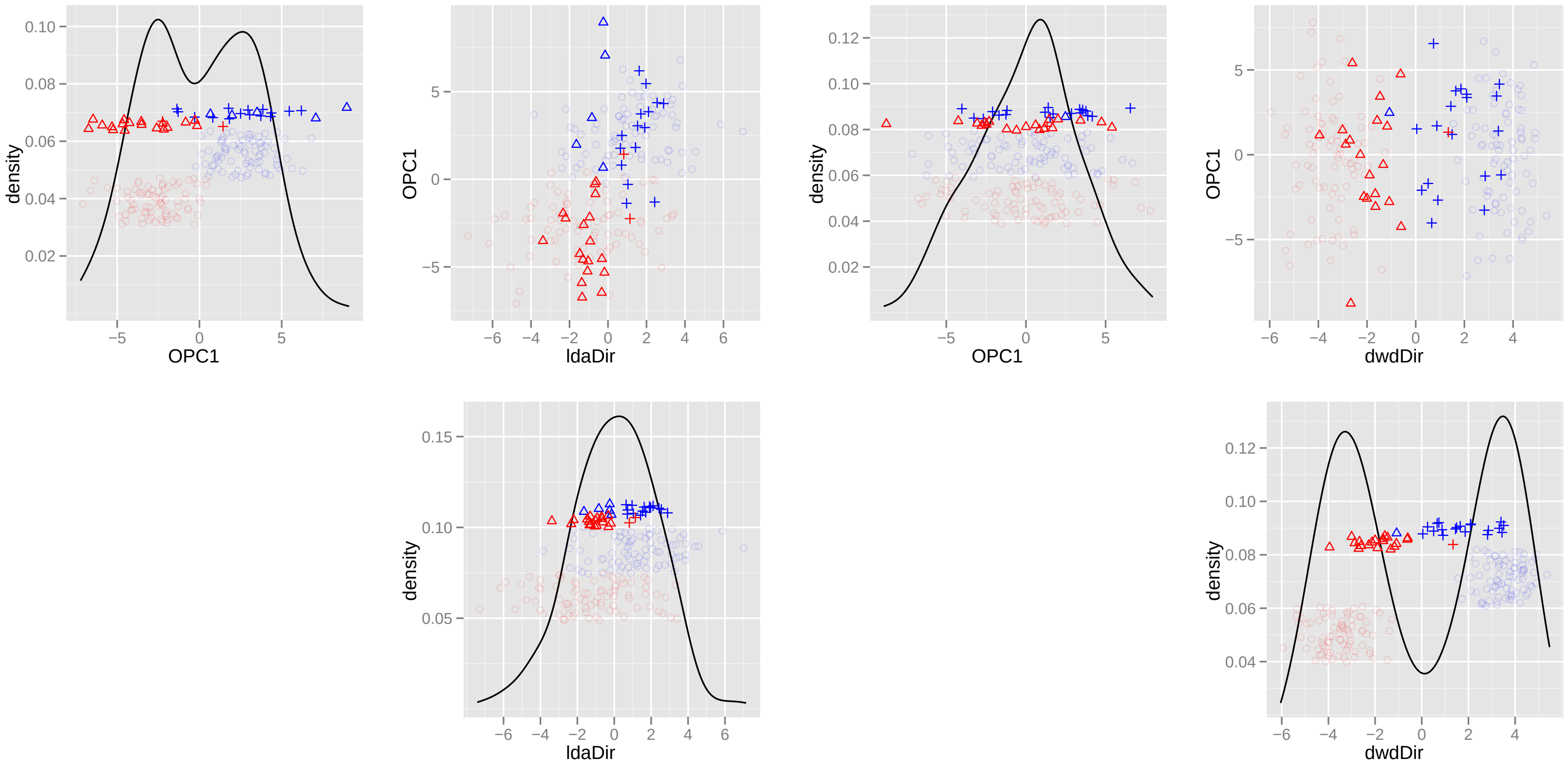

Figure 9: The left shows LDA prediction. The right shows Hanwen's DWD prediction. The color indicates the true classification. The lighter colored points (circular) are from the training data, and the rest from the test data. The symbol of the points from the test data indicates the predicted class. The corss symbol corresponds to the blue group, and the triangle correspondsto the red group.