A New Data Compression Method for Classification Analysis

Barry Wise
Manny Palacios
Donal O'Sullivan
Eigenvector Research, Inc

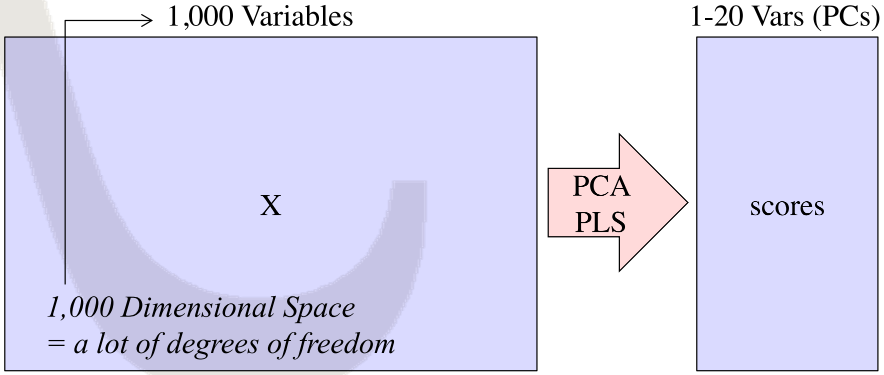
EIGENVECTOR RESEARCH INCORPORATED

“One-against-all” (OAA) PLSDA Compression

- Describe OAA-PLSDA compression for classification analyses
- Evaluate on 4 diverse datasets using SVM and XGB classification analysis
- Compare results using OAA-PLSDA compression against
  - Standard PLSDA compression or
  - no compression

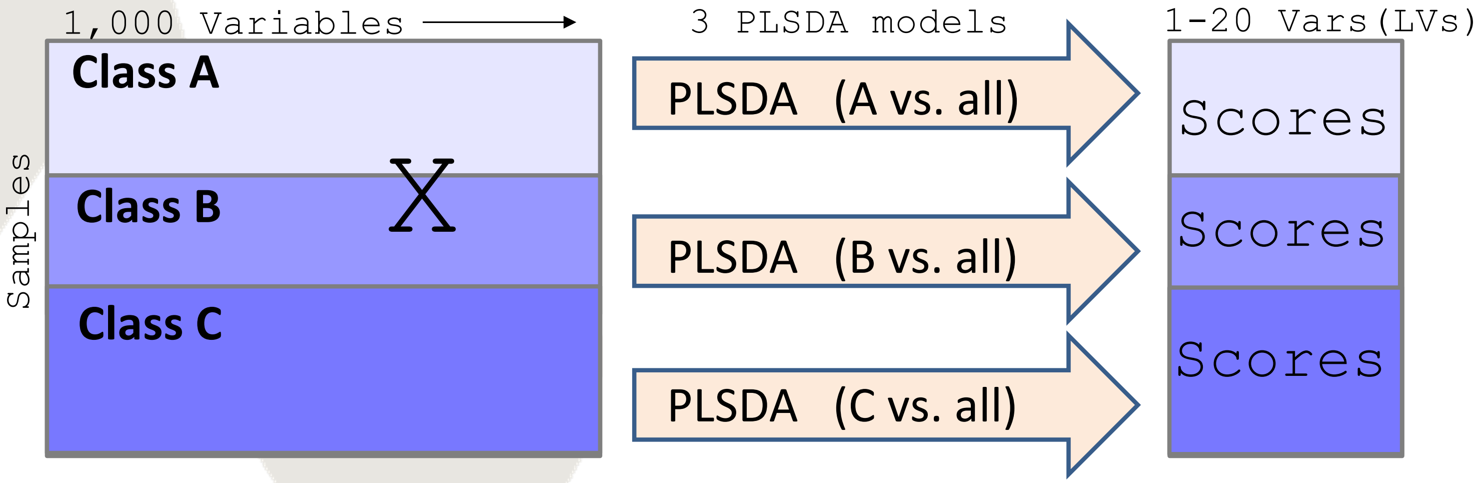
Compression

- **X-block compression:** Data compression performed on X-block prior to calculating or applying the model.
- **Compression type:** 'pca' uses a simple PCA model to compress the information. 'pls' uses a PLS or PLSDA model. Compression can make models more stable and less prone to overfitting, and faster to calculate.



“One-against-all” OAA-PLSDA Compression

Is it possible to get a better compression model for classification data by using a PLSDA model for each class, instead of using one overall PLSDA model?



“One-against-all” OAA-PLSDA Compression

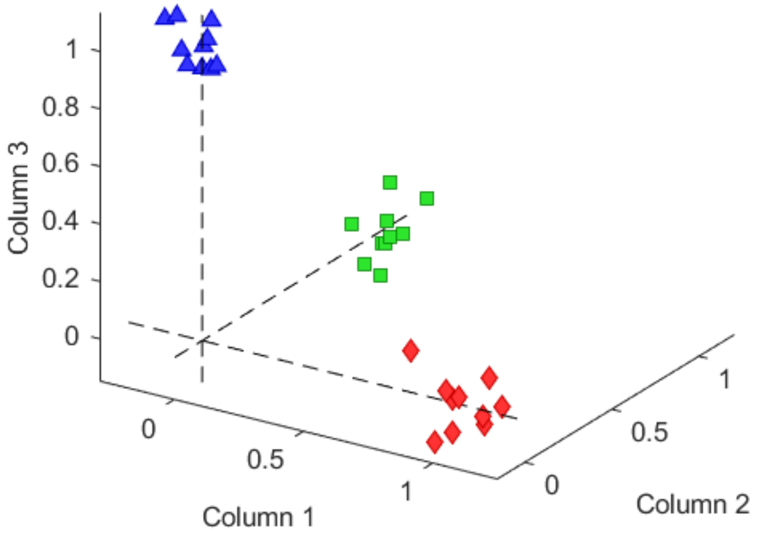
- Use one-against-all PLSDA compression models for each class
- Build a PLSDA model for each class against all others ("one-against-all")
- Use the scores and/or predictions from these Nclass models
- Data size = (m, n) compresses to size = (m, (ncomp+1)*nclass) if scores and predictions are used, and ncomp LVs are used

Test OAA-PLSDA compression used with SVM and XGB Discriminant Analysis

- Compare SVMDA and XGBoostDA prediction performance
- Compare OAA-PLSDA compression against PLS compression and No-Compression
- Use 4 datasets:
  1. Synthetic 3-class dataset (linearly separable)
  2. Synthetic 3-class dataset (not linearly separable)
  3. Hyperspectral aerial image dataset using 3 classes
  4. Large LIBS dataset using 5 classes
- Compare results using Misclassification error rate for each class, or the proportion of samples which were incorrectly classified (FP +FN)/N

1. Separable Synthetic Dataset

3 classes. Data size = (3000, 600)
Data are linearly separable (except for noise)



Data values:
Class 1: Samples 1-1000 have
variables 1:200 = 1, others = 0
Class 2: Samples 1001-2000 have
variables 201:400 = 1, others = 0
Class 3: Samples 2001-3000 have
variables 401:600 = 1, others = 0

Plus Gaussian distributed noise centered on origin added to all variables

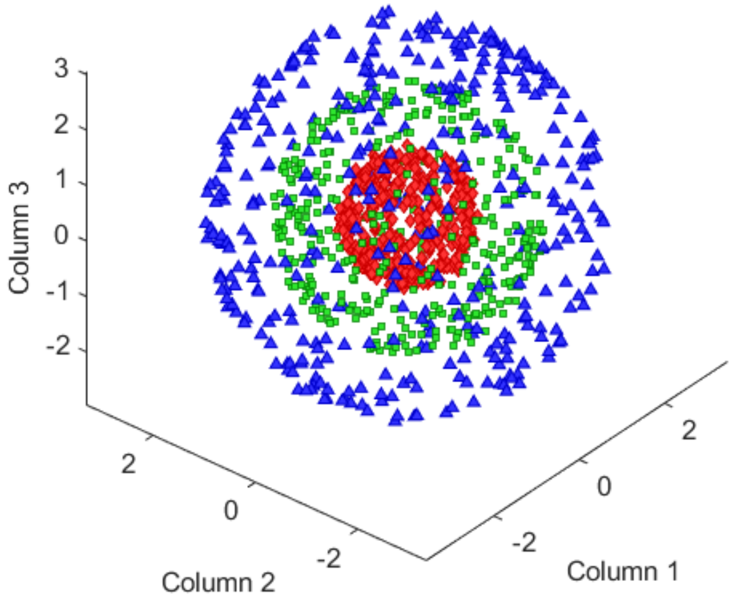
EIGENVECTOR RESEARCH INCORPORATED

2. Non-separable Synthetic Dataset

3 classes. Data size = (1200, 100)

Data are not linearly separable

All samples:
3 variables used for class shells
97 variables are Gaussian noise



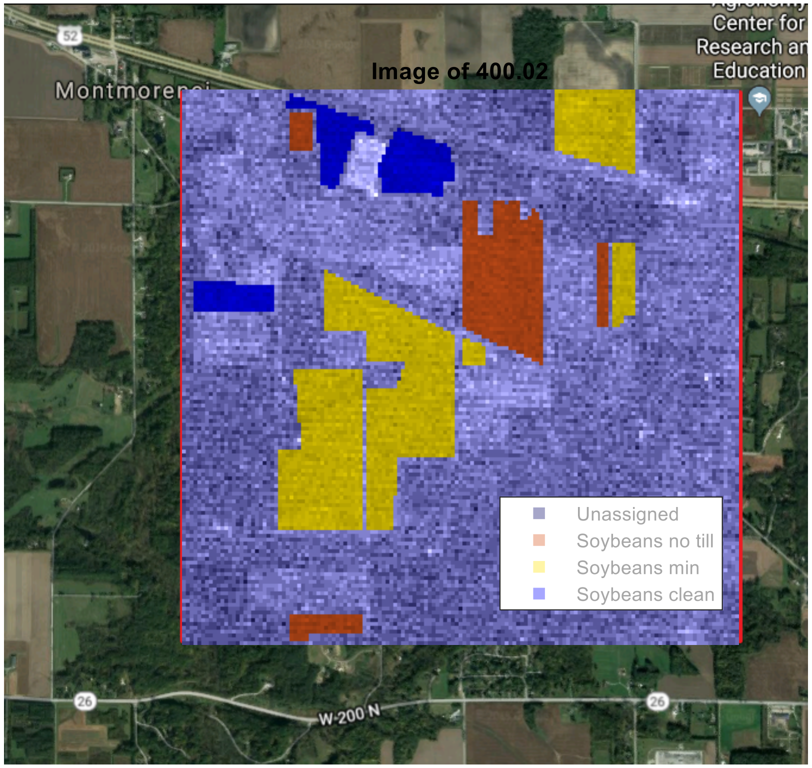
EIGENVECTOR RESEARCH INCORPORATED

3. Aerial Hyperspectral Image Dataset

3 classes. Data size = (3341, 220)

Hyperspectal image of mixed farmland. Image has 220 spectral channels

Using 3341 pixels from Soy fields, which are 3 types: “No till”, “Min till” and “Clean”

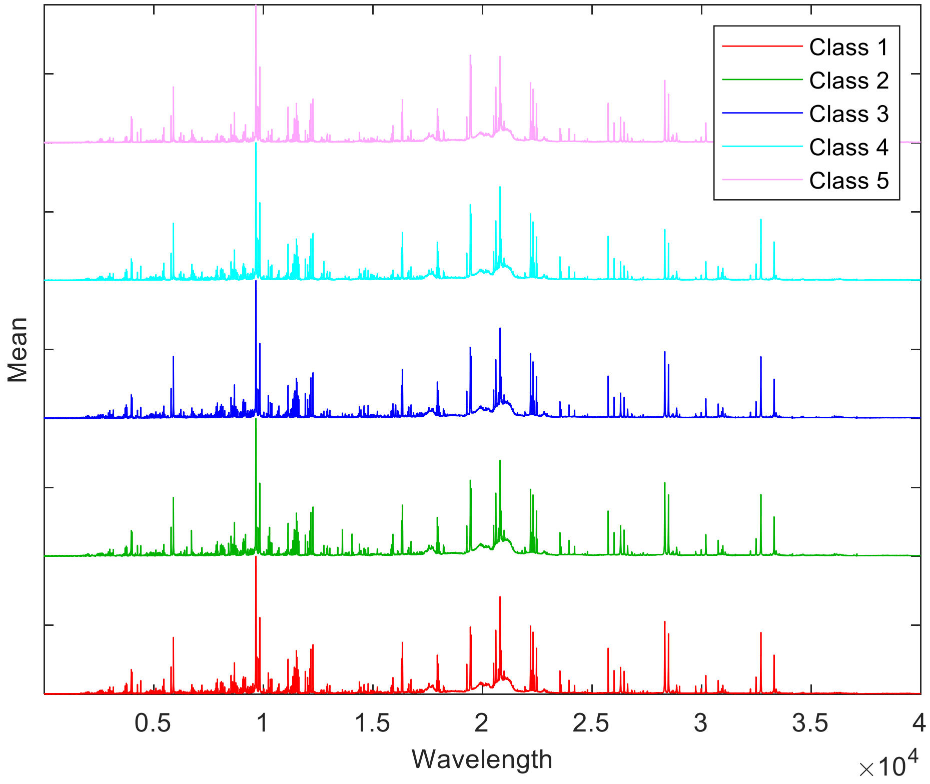


EIGENVECTOR RESEARCH INCORPORATED

4. LIBS Dataset

5 classes. Data size = (1050,40002)

Figure shows the 5 classes offset for visibility

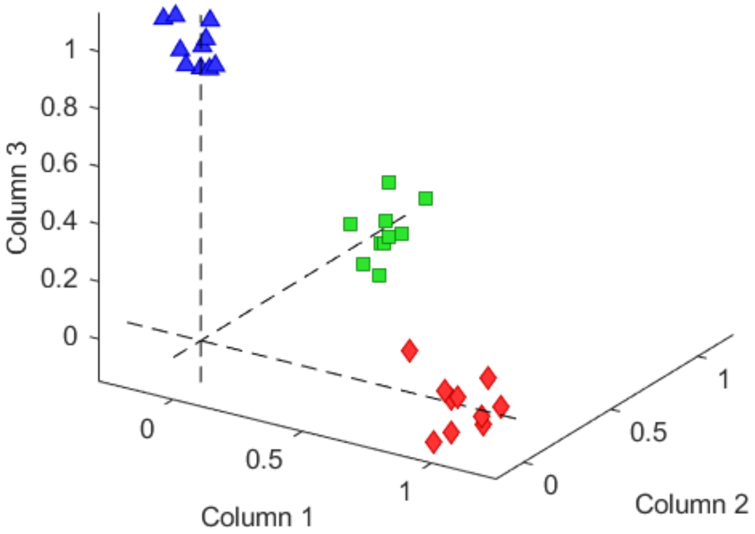


1. Separable Synthetic Dataset

EIGENVECTOR RESEARCH INCORPORATED

1. Separable Synthetic Dataset

3 classes. Data size = (3000, 600)
Data are linearly separable (except for noise)



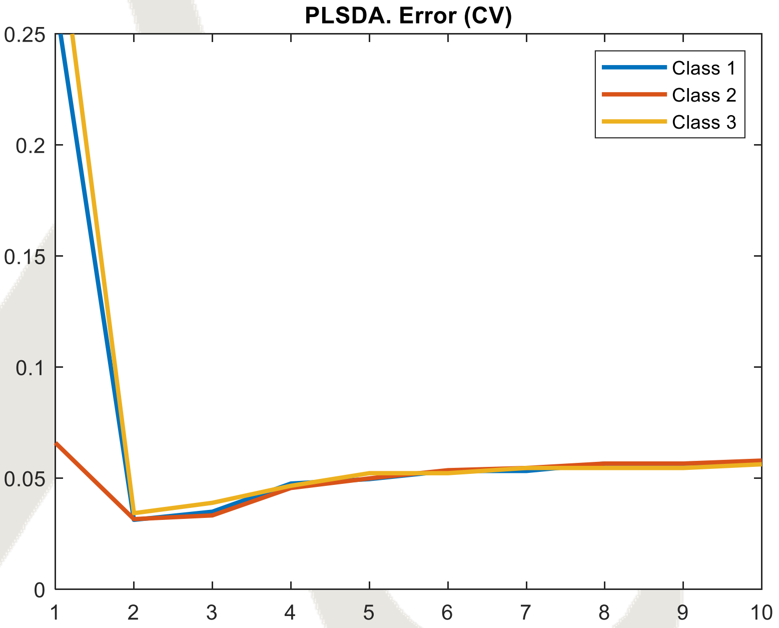
Data values:

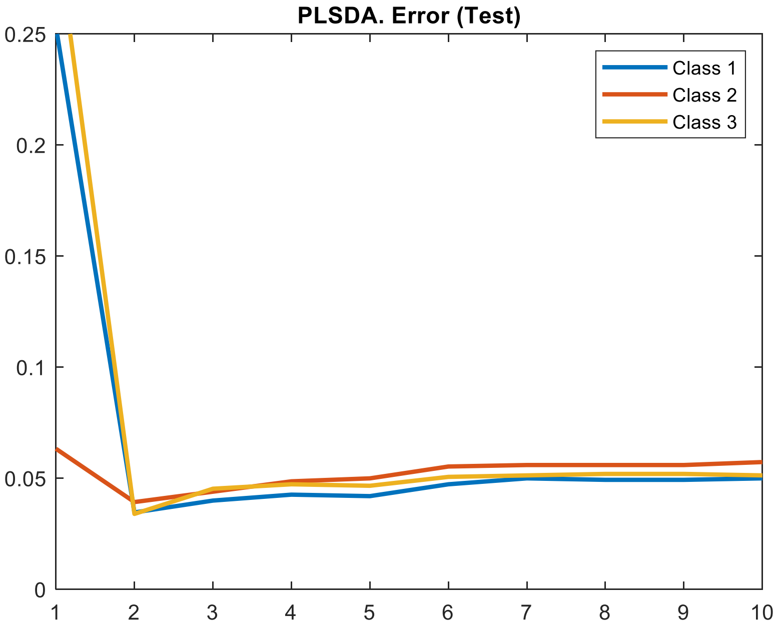
Class 1: Samples 1-1000 have variables 1:200 = 1, others = 0

Class 2: Samples 1001-2000 have variables 201:400 = 1, others = 0

Class 3: Samples 2001-3000 have variables 401:600 = 1, others = 0

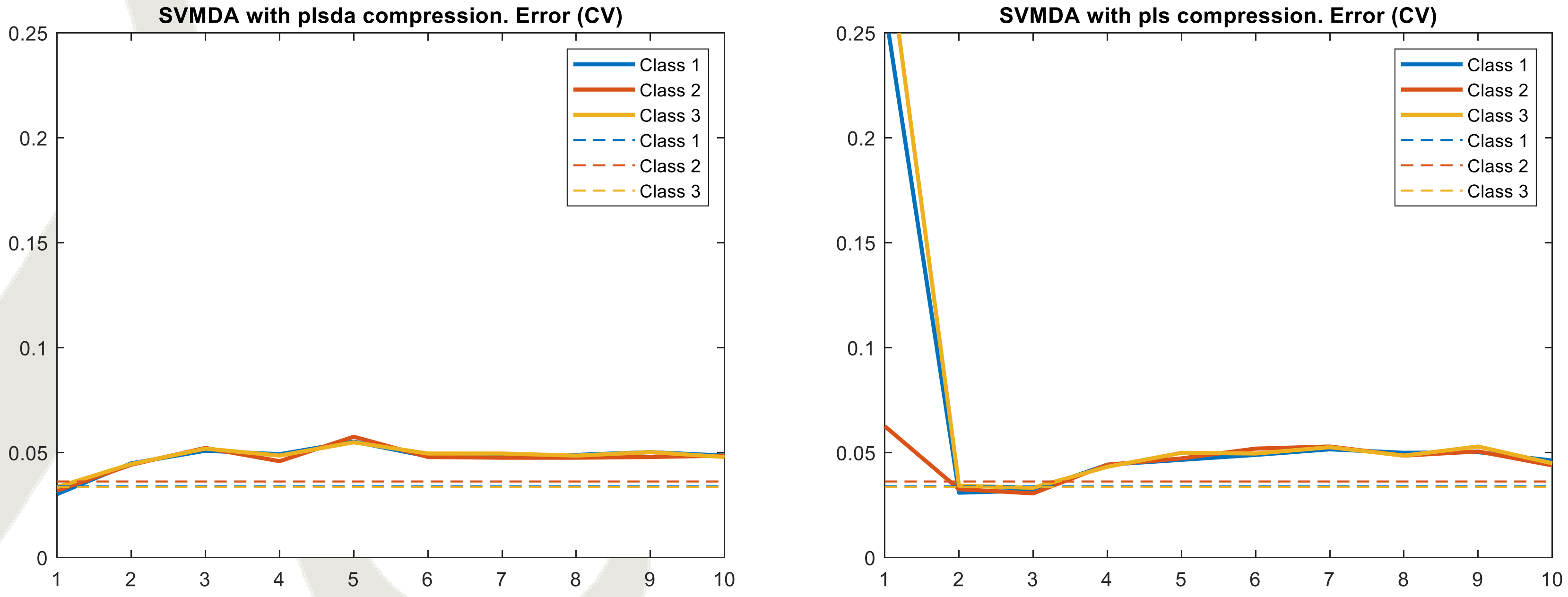
Plus Gaussian distributed noise centered on origin added to all variables





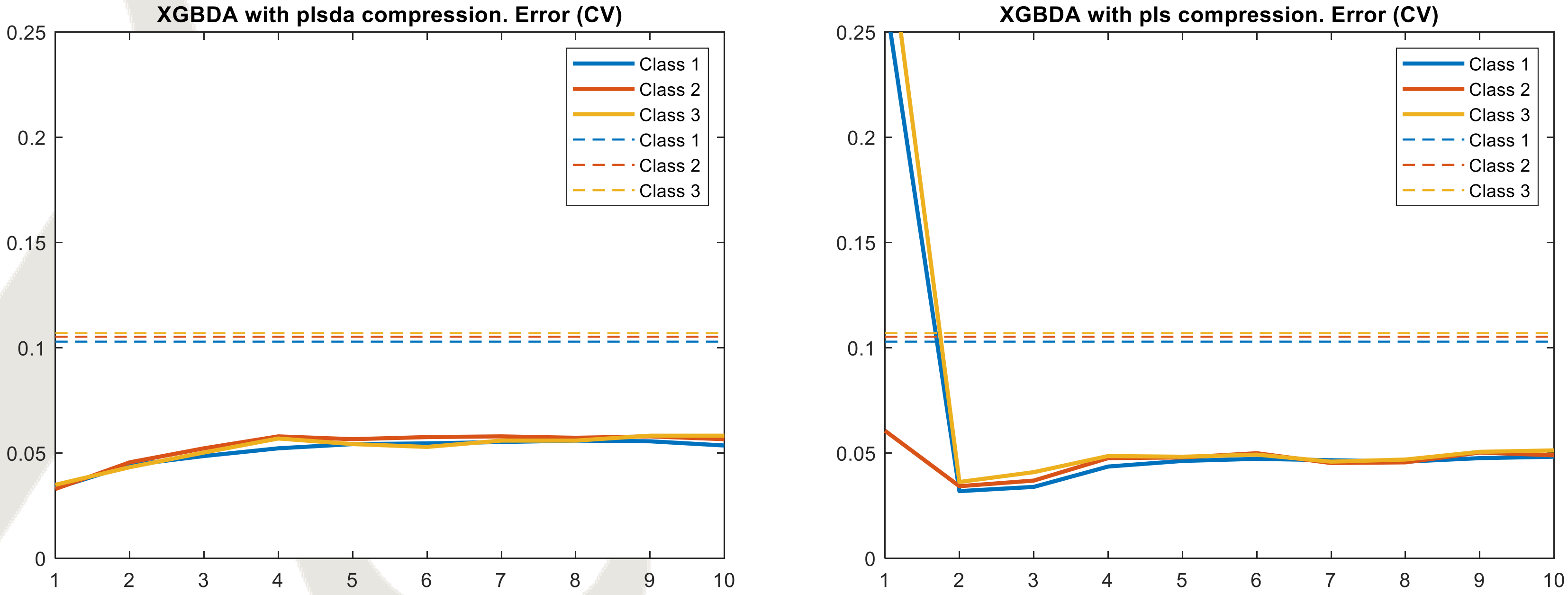
EIGENVECTOR
RESEARCH INCORPORATED

Separable Synthetic Dataset: SVMDA Classification Error



Dashed lines show the error for the no-compression case

Separable Synthetic Dataset: XGBDA Classification Error



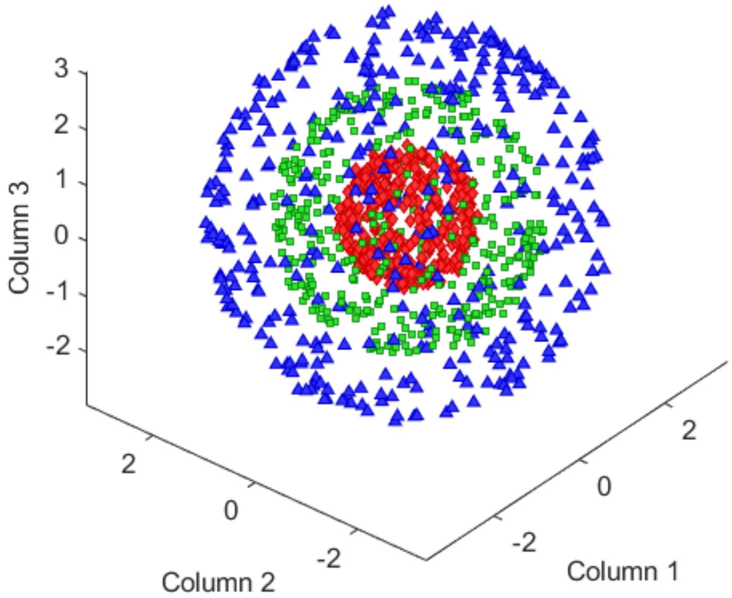
Dashed lines show the error for the no-compression case

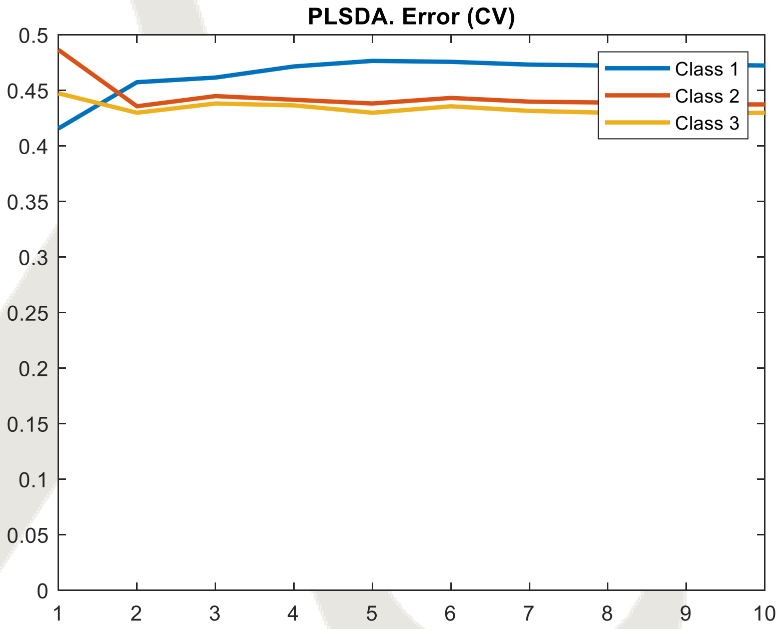
2. Non-separable Synthetic Dataset

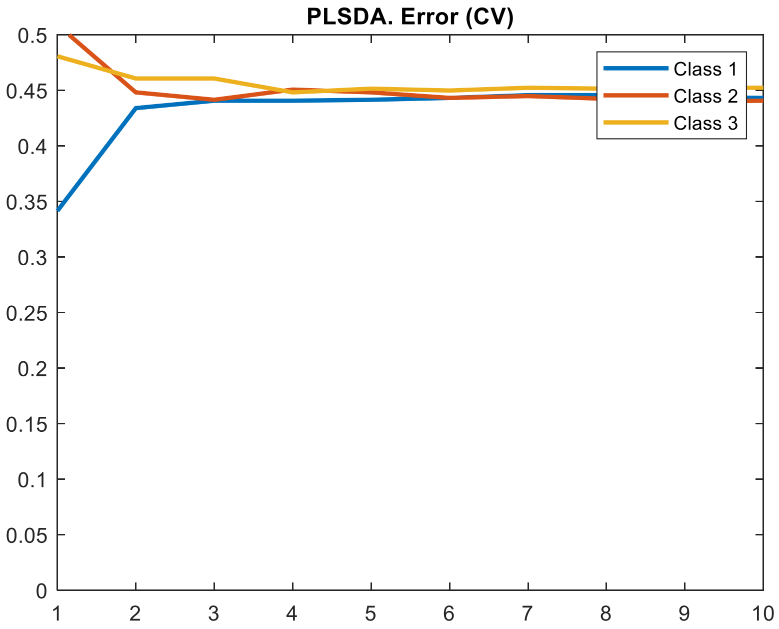
3 classes. Data size = (1200, 100)

Data are not linearly separable

All samples:
3 variables used for class shells
97 variables are Gaussian noise

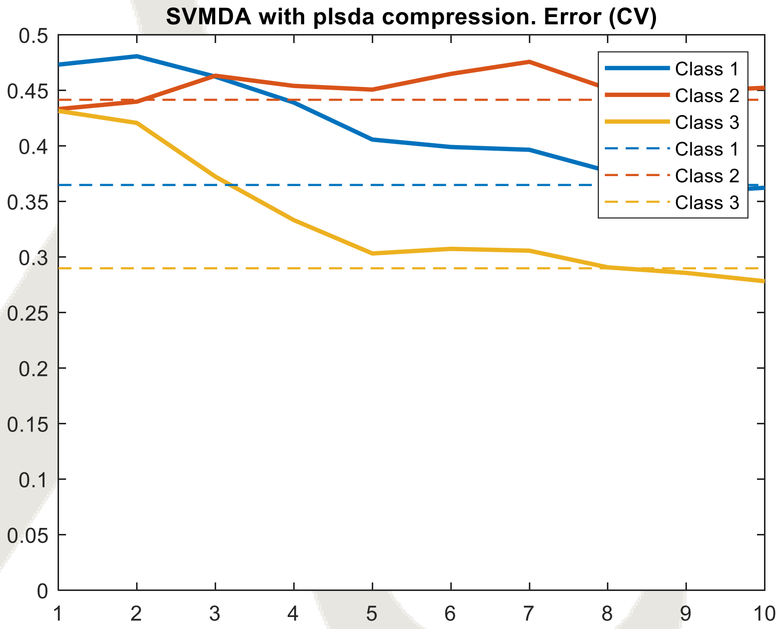


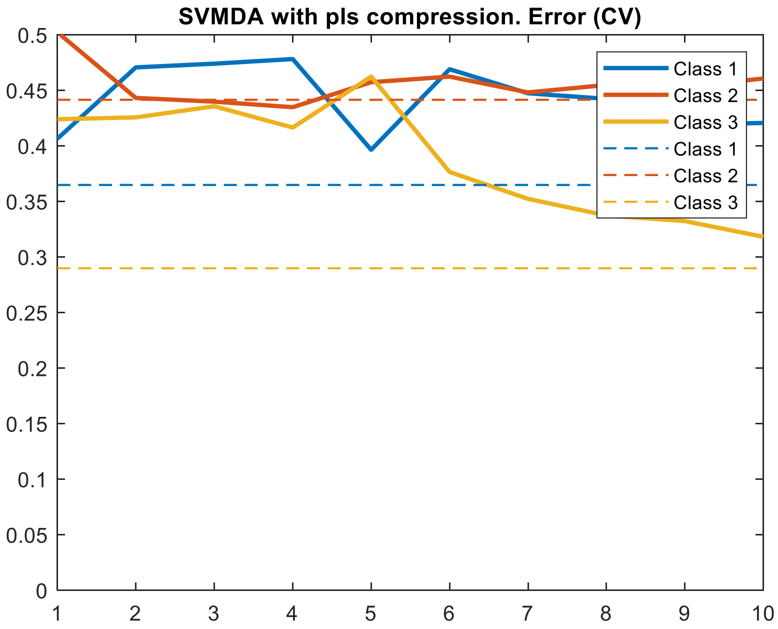

PLSDA. Error (CV)


PLSDA. Error (CV)

EIGENVECTOR
RESEARCH INCORPORATED

Non-separable Synthetic Dataset: SVMDA Classification Error

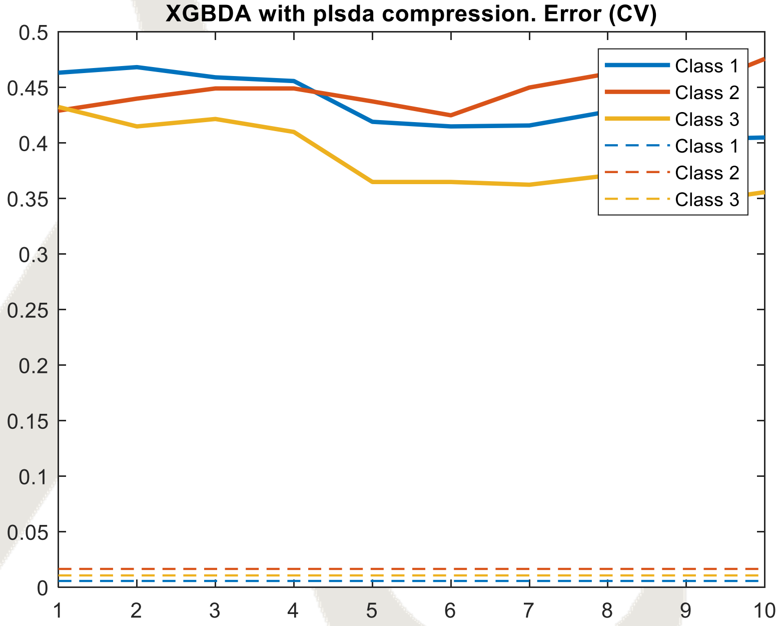


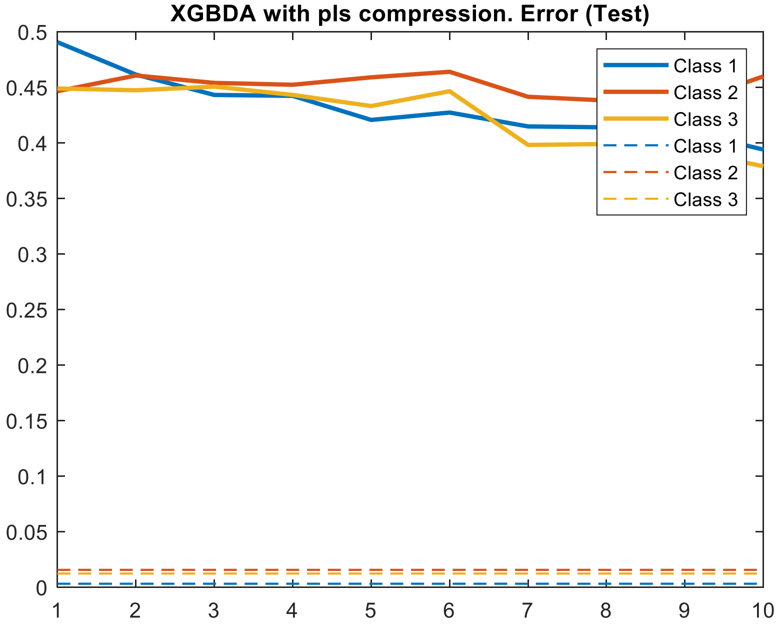


Dashed lines show the error for the no-compression case

EIGENVECTOR RESEARCH INCORPORATED

Non-separable Synthetic Dataset: XGBDA Classification Error





Dashed lines show the error for the no-compression case

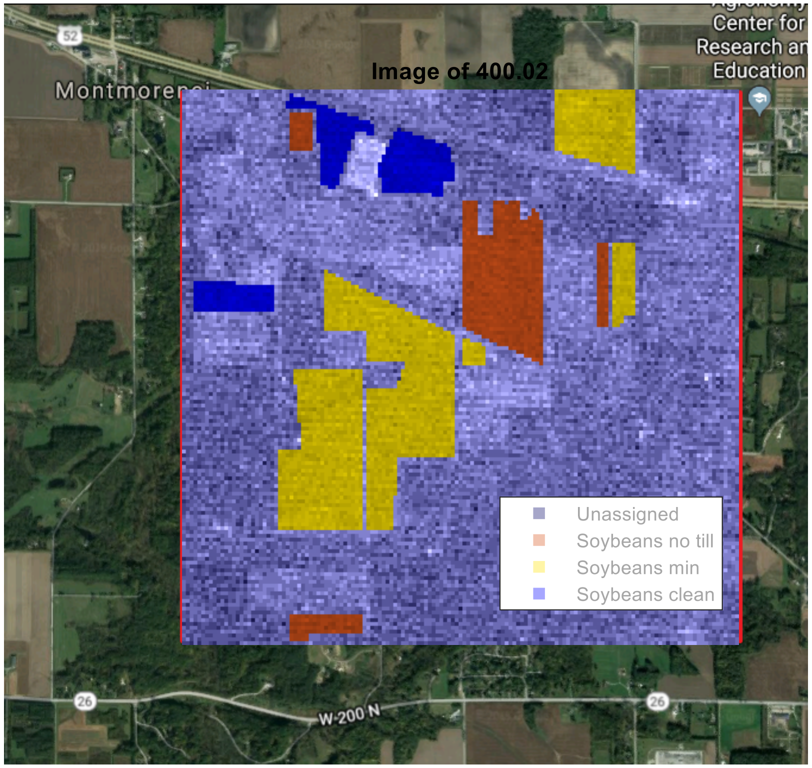
EIGENVECTOR RESEARCH INCORPORATED

3. Aerial Hyperspectral Image Dataset

3 classes. Data size = (3341, 220)

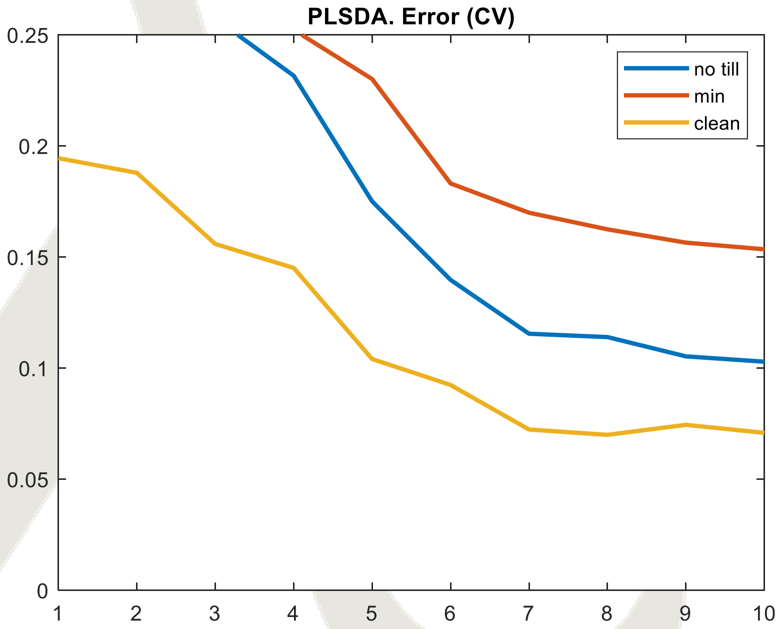
Hyperspectal image of mixed farmland. Image has 220 spectral channels

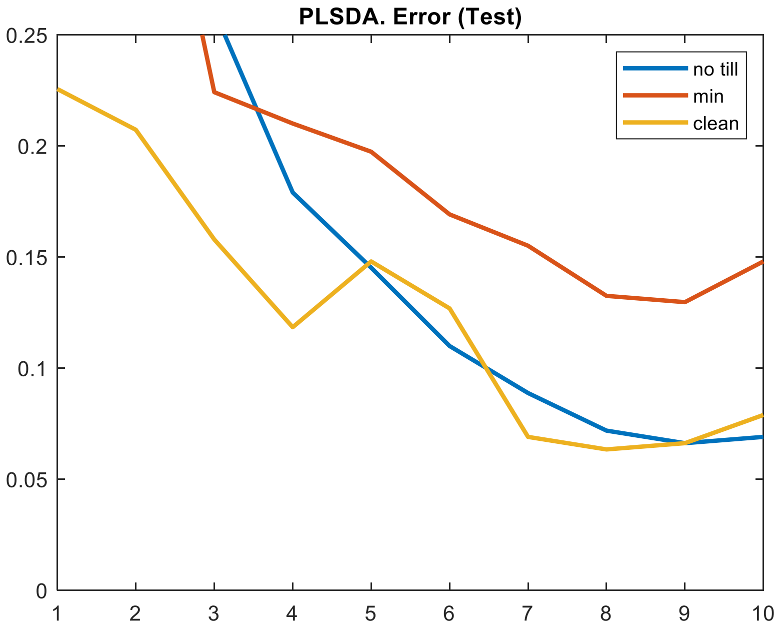
Using 3341 pixels from Soy fields, which are 3 types: “No till”, “Min till” and “Clean”



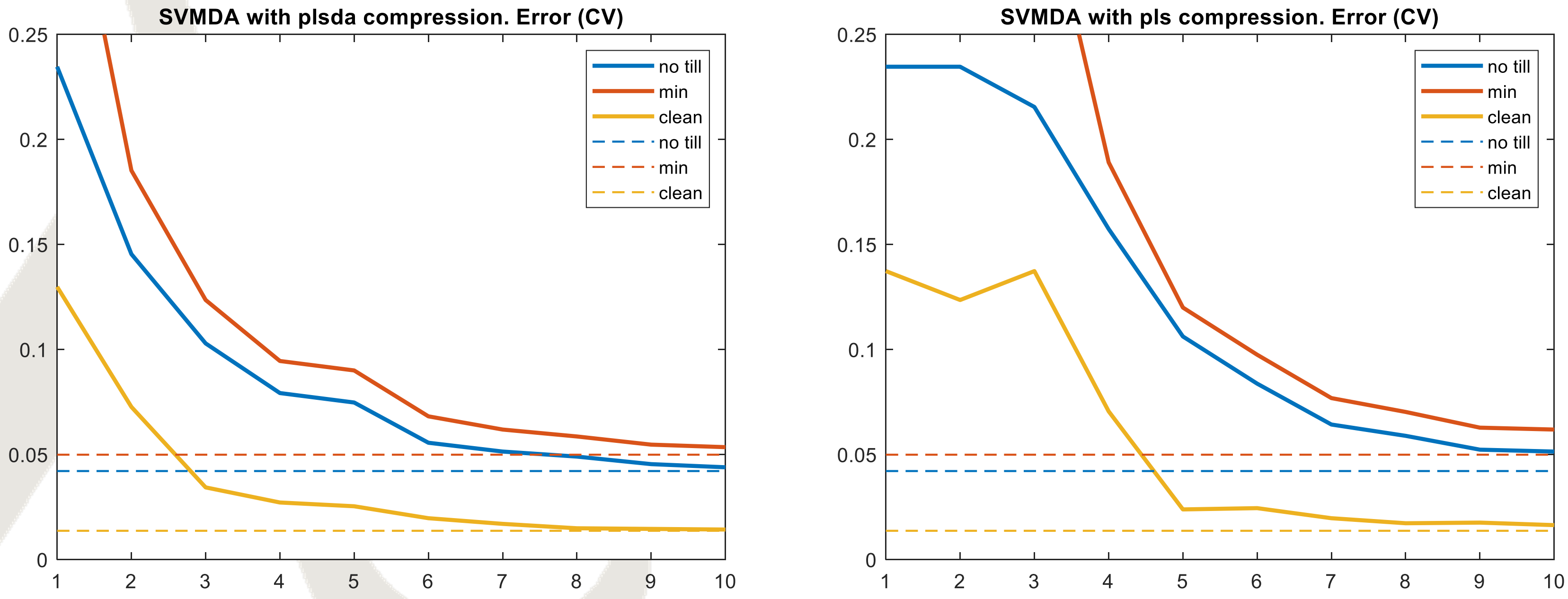
EIGENVECTOR RESEARCH INCORPORATED

Aerial Hyperspectal Image


PLSDA. Error (CV)

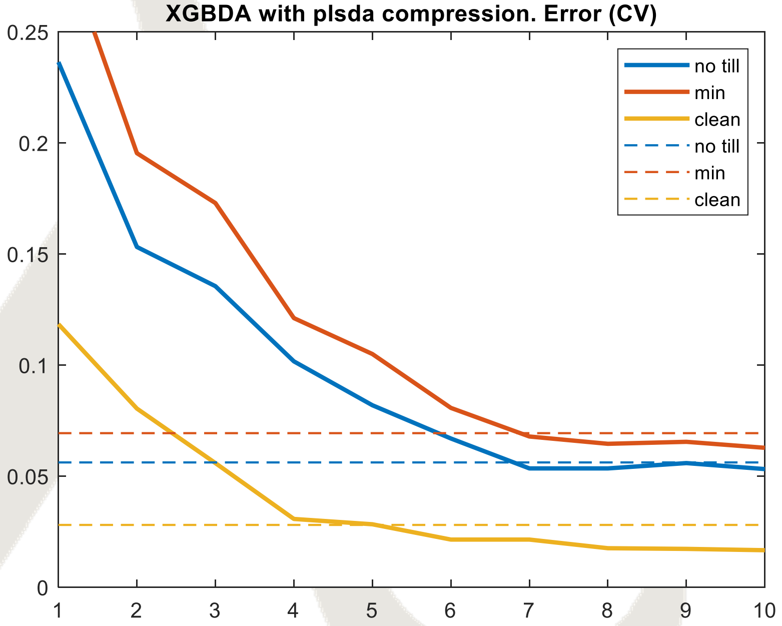

PLSDA. Error (Test)

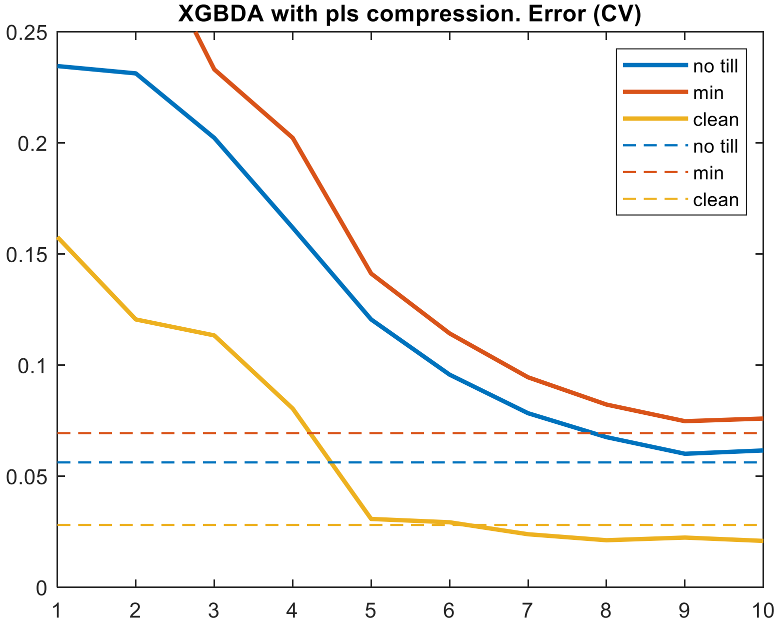
Aerial Hyperspectal Image: SVMDA Classification Error



Dashed lines show the error for the no-compression case

Aerial Hyperspectal Image: XGBDA Classification Error





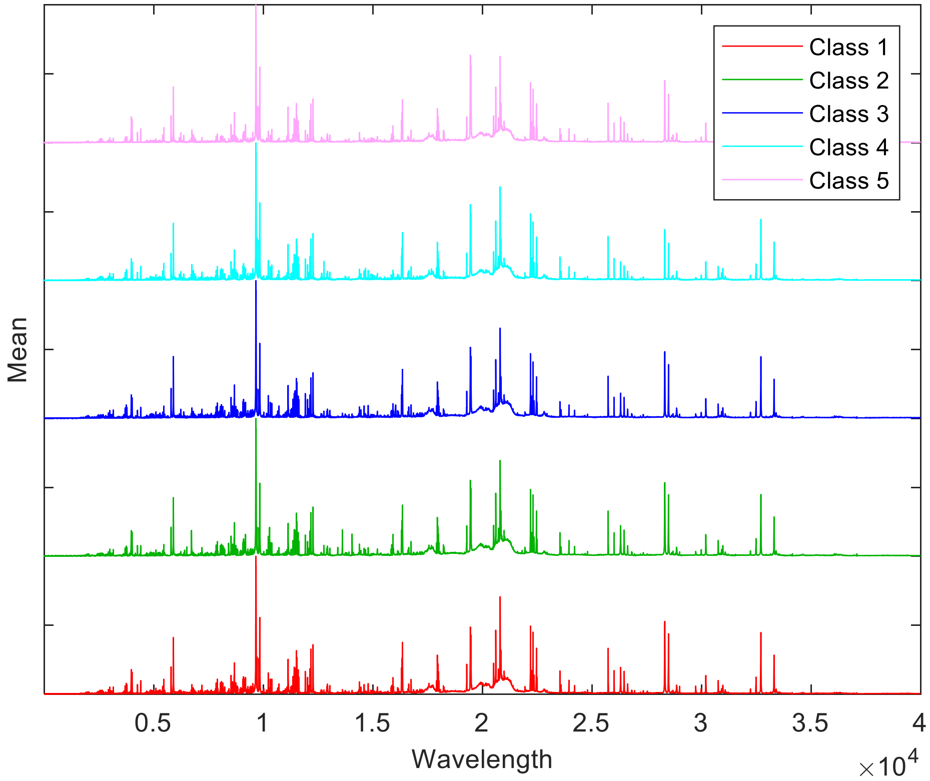
Dashed lines show the error for the no-compression case

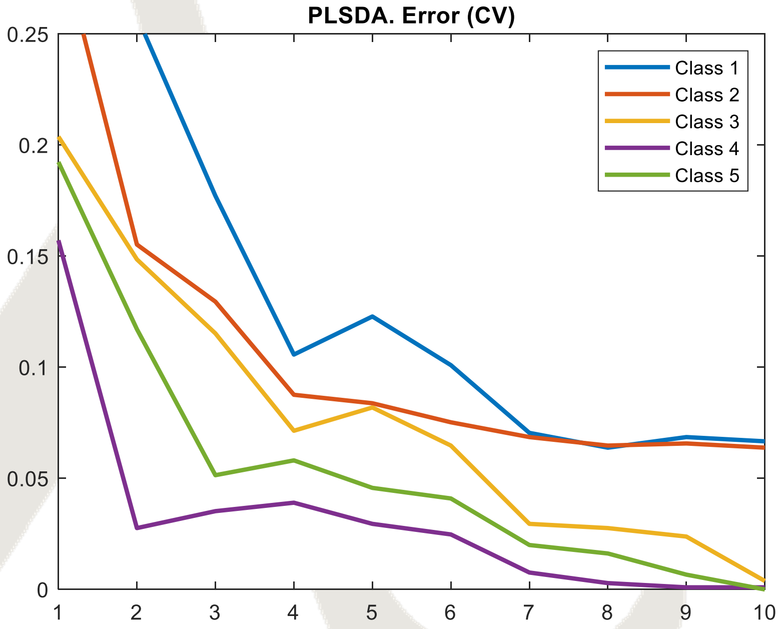
EIGENVECTOR
RESEARCH INCORPORATED

4. LIBS Dataset

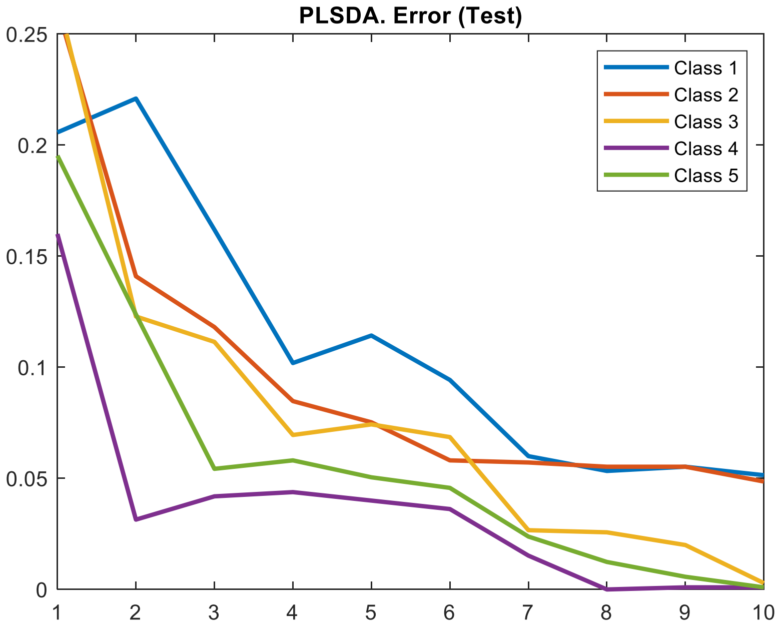
5 classes. Data size = (1050,40002)

Figure shows the 5 classes offset for visibility





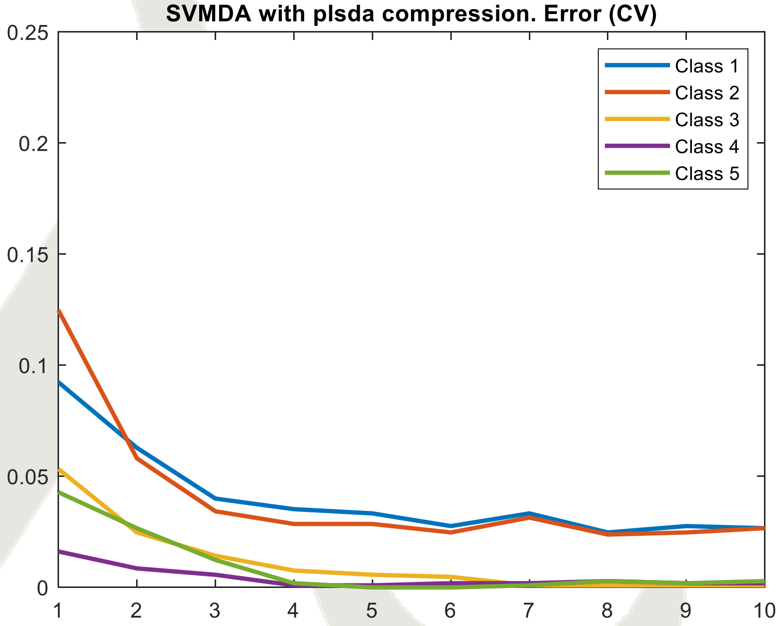
PLSDA. Error (CV)

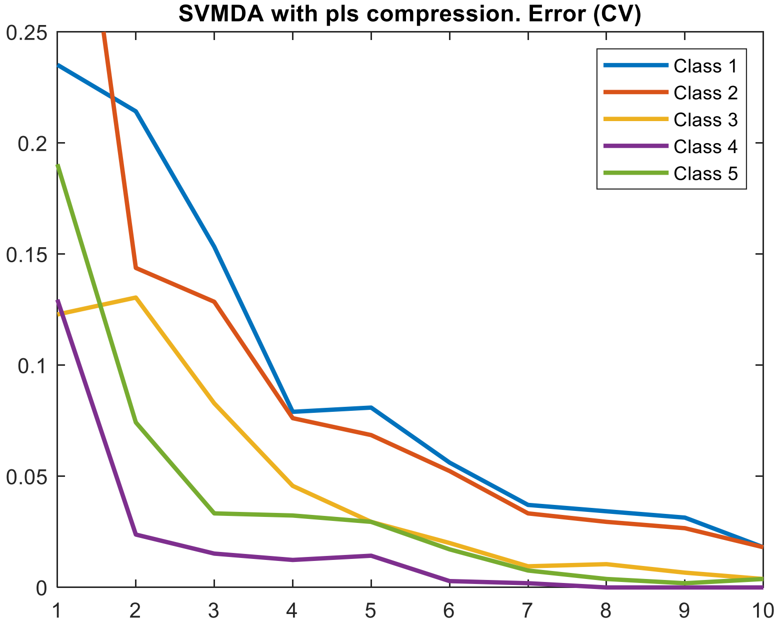


PLSDA. Error (Test)

EIGENVECTOR
RESEARCH INCORPORATED

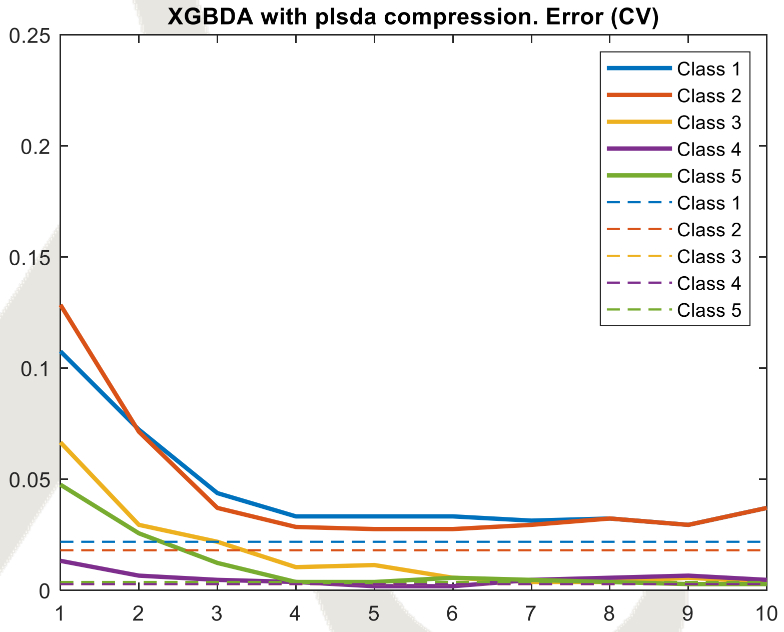
LIBS Dataset: SVMDA Classification Error

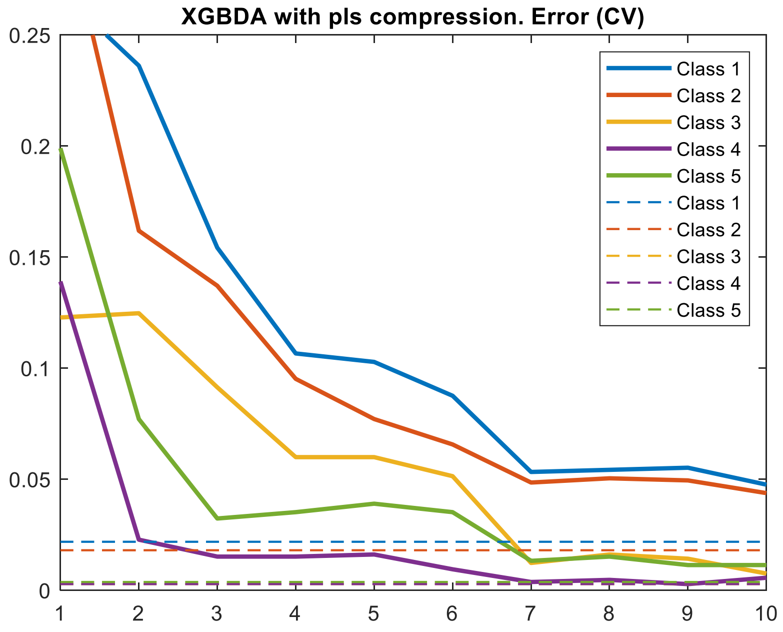




EIGENVECTOR RESEARCH INCORPORATED

LIBS Dataset: XGBDA Classification Error


XGBDA with plsda compression. Error (CV)


XGBDA with pls compression. Error (CV)

Dashed lines show the error for the no-compression case

EIGENVECTOR RESEARCH INCORPORATED

Conclusions

- OAA-PLSDA performs similarly to PLSDA compression for classification using SVMDA or XGBoostDA but appears to be more concise, getting better results when using low number of compression latent variables.

- Compression using PLSDA or OAA-PLSDA will not capture some nonlinearity – as shown by the non-separable case. Using such compression before SVMDA or XGBoostDA will not help

EIGENVECTOR RESEARCH INCORPORATED