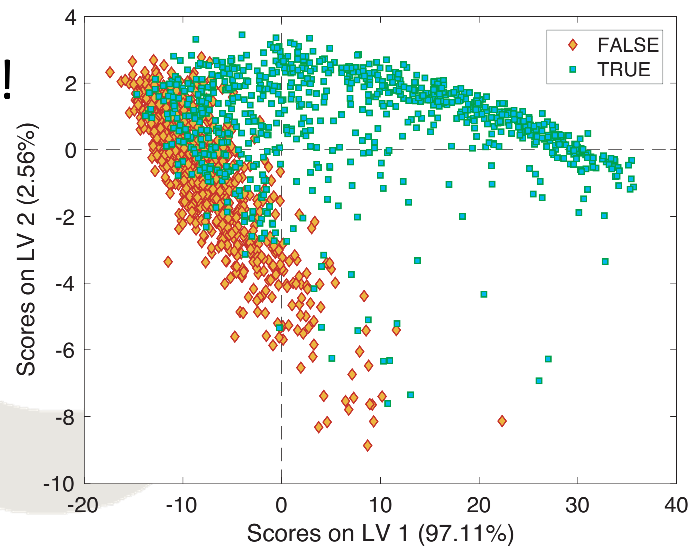
A Comparison of ANNs, SVMs & XGBoost on some Challenging Classification Problems

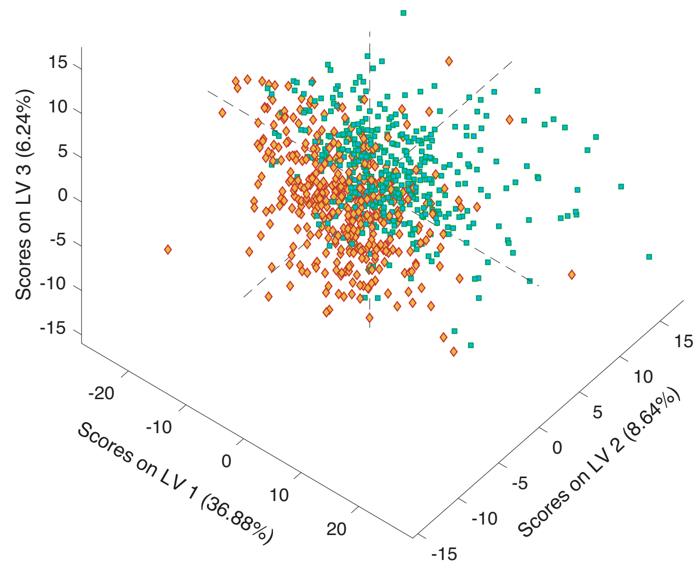
Barry M. Wise, Donal O’Sullivan and
Manuel A. Palacios

Eigenvector Research, Inc.

Challenging?

- What do you mean by challenging?
  - Classification problems where we're expecting 80-90% correct (not close to 100%!)
- Why?
  - Only kind we get!





EIGENVECTOR RESEARCH INCORPORATED

Outline

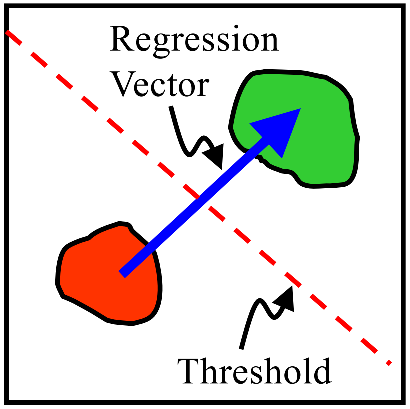
- Classification methods
- The data sets
- Results
- Conclusions

Classification Methods

- PLS-DA Partial Least Squares Discriminant Analysis
- ANN Artificial Neural Network
- SVM Support Vector Machine
- XGBoost Boosted Trees Classification

Partial Least Squares Discriminant Analysis (PLS-DA)

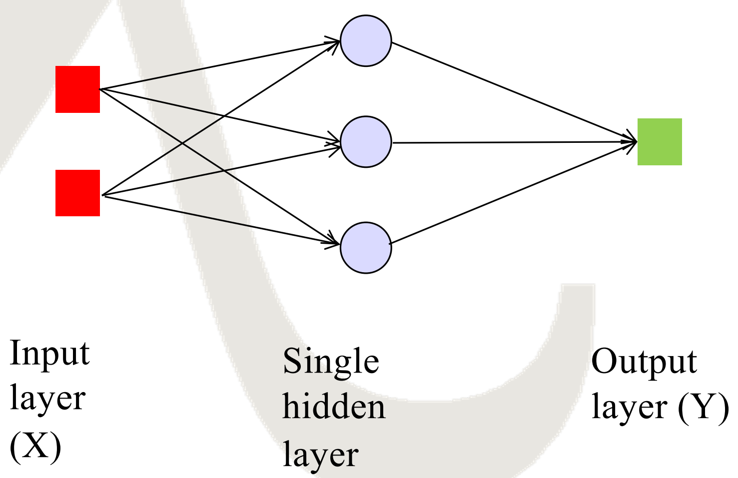
- A true workhorse of classification methods!
- Use logicals (0,1) in Y-block to indicate if sample belongs to a class or not ➔ dummy variables
- Develop PLS model to predict class block
- Thresholds set between 0 and 1 to indicate if new samples are a member of each class...
  - Can use Bayes theorem to set threshold and include prior probability and set costs



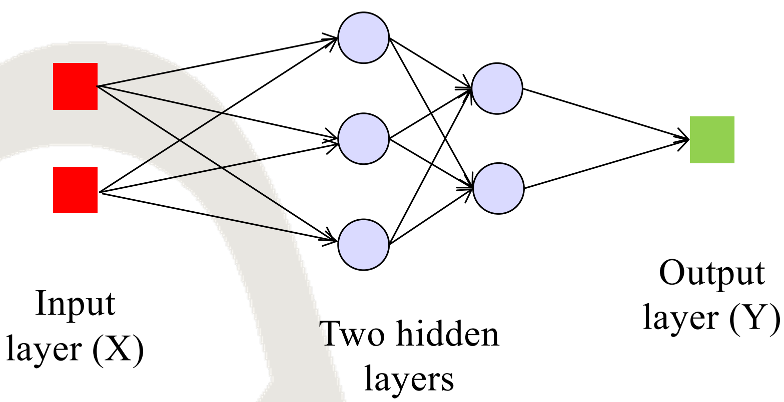
EIGENVECTOR RESEARCH INCORPORATED

Artificial Neural Networks

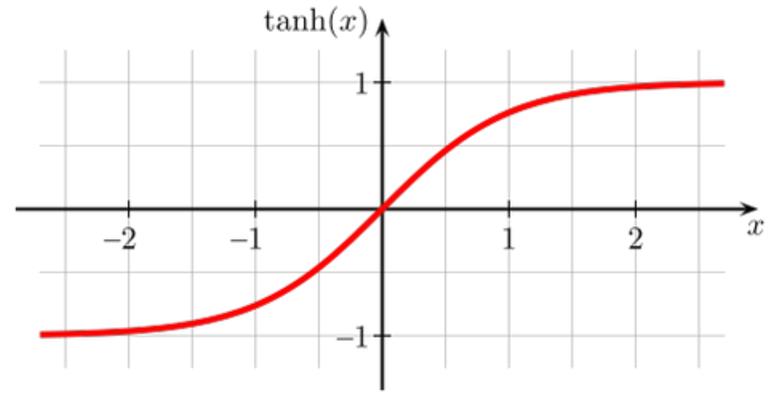
- Artificial Neural Network (ANN) is a non-linear regression method.
- X data are presented to the ANN in the input layer. A simple single hidden-layer example:



If the input to a neuron is strong enough the neuron is activated and it affects downstream connected neurons



ANN



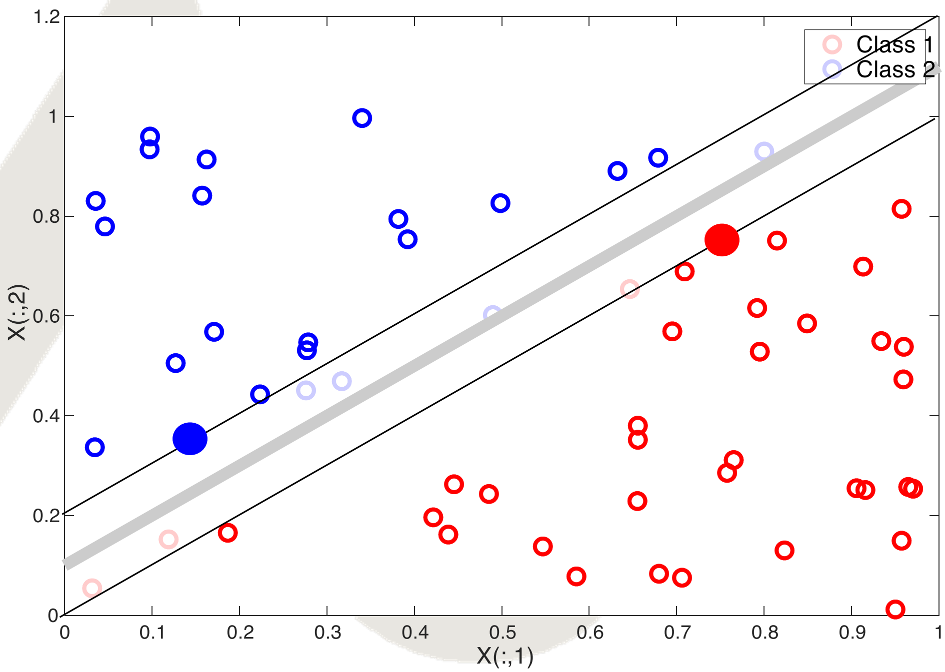
- ANNs defined by
  - Layers and nodes in each layer and their connections.
  - Weights: weight associated with each synapse, or node-pair.
  - Activation function converts node's weighted input to its output, usually step-like such as tanh.
- For classification predict logicals as with PLS-DA
- Fit via least squares optimization

Support Vector Machines

Support Vector Machines (SVMs) are a set of related supervised learning techniques for **classification** and **regression** which became popular over the past decade.

SVM Classification

SVMs finds the optimal separating margin between each pair of classes.



$\min(\mathbf{w}^T\mathbf{w})$ subject to $y_i([\mathbf{w}^T\mathbf{x}_i+b])\geq 1$

Support vectors = the samples where the equality holds. The ones further out don't matter, once **w** and *b* are found

EIGENVECTOR RESEARCH INCORPORATED

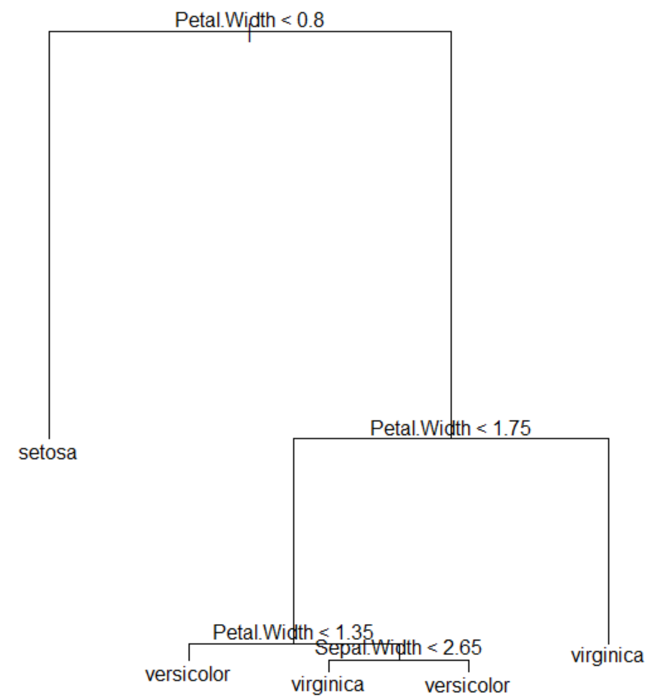
SVM Parameters

- SVM classification involves defining parameters (**cost**, **gamma**).
  - Cost: (0 – infinity). When high, allow less misclassification but could cause overfitting.
  - gamma: (0 – infinity). Low, linear; high local and nonlinear
- The SVM function selects automatically by default using cross-validation.

Classification and Regression Trees

- **Regression Trees:**
  - Algorithm picks splitting variables & split points.
  - Minimizes sum of squares of y – f(x).
  - Test each variable and split point picking the one which gives min sum of squares error.
  - Prediction value is given by the leaf value.
- **Classification Trees:**
  - Instead of squared error uses a measure of impurity, misclassification error, Gini index, cross-entropy, to select the best binary decision.

Classification Tree using "iris" dataset



Boosting

- Classification and Regression trees have many advantages but not great accuracy, hence Boosting is used
- The motivation for boosting is to combine the outputs of many "weak" classifiers to produce a powerful classifier
- Additive boosting (Adaboost) binary classification, increases weights of observations which are misclassified and classifies again, producing a sequence of classifiers.
- Gradient Boosting applied to decision trees creates new trees which best reduce an error loss function by using gradient descent.

EIGENVECTOR RESEARCH INCORPORATED

Why XGBoost?

XGBoost is an open-source implementation of gradient boosted decision trees

- XGBoost is a freely available (Apache License 2.0)

  http://dmlc.cs.washington.edu/xgboost.html

- Released in 2014, by UW, it is written in C++ with interfaces for many languages including Python, R, **Java**…
- Currently very popular with machine learning data analysts
- It is accuracy, fast, scales well with computing resources,…

EIGENVECTOR RESEARCH INCORPORATED

…and XGBoost has **Hype!**

If linear regression was a Toyota Camry, then gradient boosting would be a UH-60 Blackhawk Helicopter. A particular implementation of gradient boosting, XGBoost, is consistently used to win machine learning competitions on Kaggle. Unfortunately many practitioners (including my former self) use it as a black box. It's also been butchered to death by a host of drive-by data scientists' blogs. As such, the purpose of this article is to lay the groundwork for classical gradient boosting, intuitively *and* comprehensively.

http://blog.kaggle.com/2017/01/23/a-kaggle-master-explains-gradient-boosting/

EIGENVECTOR RESEARCH INCORPORATED

Compression

- Common to use PLS or PCA for compression in front of ANNs, SVMs, and XGBoost
- Full rank
  – Reduces problem size, speeds computation
- Reduced rank
  – Improves parsimony, possible better results

Data Sets

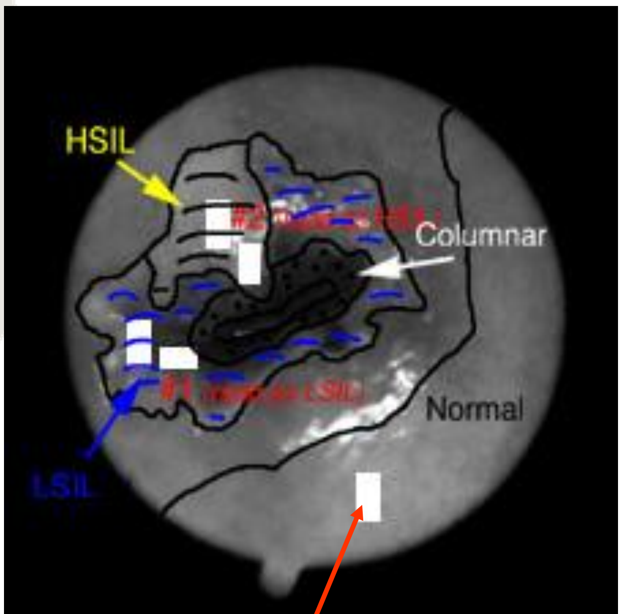
- Cervical Cancer Detection
- Breast Cancer Detection
- Infectious Disease Detection
- Hyperspectral Image for Crop Classification

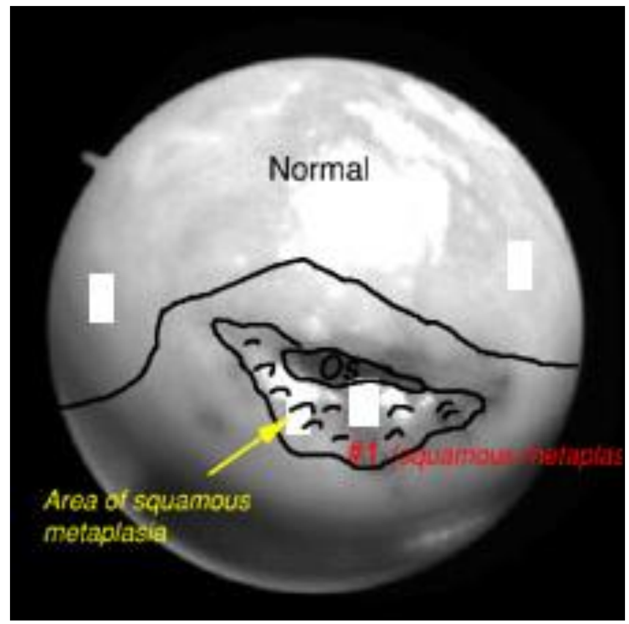
Cervical Cancer

- Pap smears credited with reducing cervical cancer mortality by detecting pre-cancerous cells, but…
- Sensitivity of Pap smears reported as 29-56%
- Abnormal Pap smear-> colposcopic examaination, but….
- Colposcopic impressions correlate with biopsies as little as 35% of the time
- Goal: develop better method to classify cervical tissue!

EIGENVECTOR RESEARCH INCORPORATED

Colposcopic Images and Biopsies



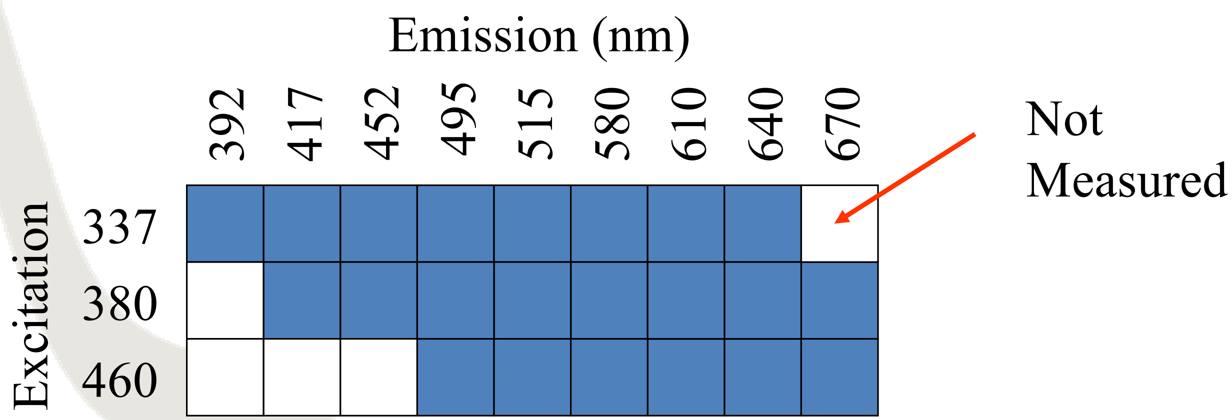


Biopsy locations

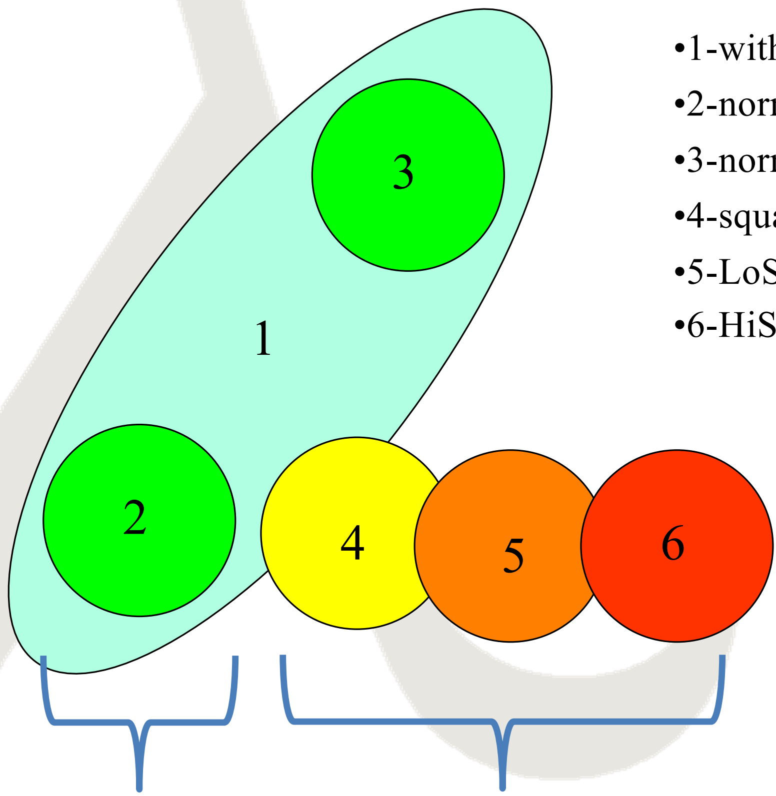
EIGENVECTOR RESEARCH INCORPORATED

Flourescence Images

- Combinations of
  - 3 excitation wavelengths
  - 9 emission wavelengths
  - 22 combinations measured



Similarity of Tissue Types

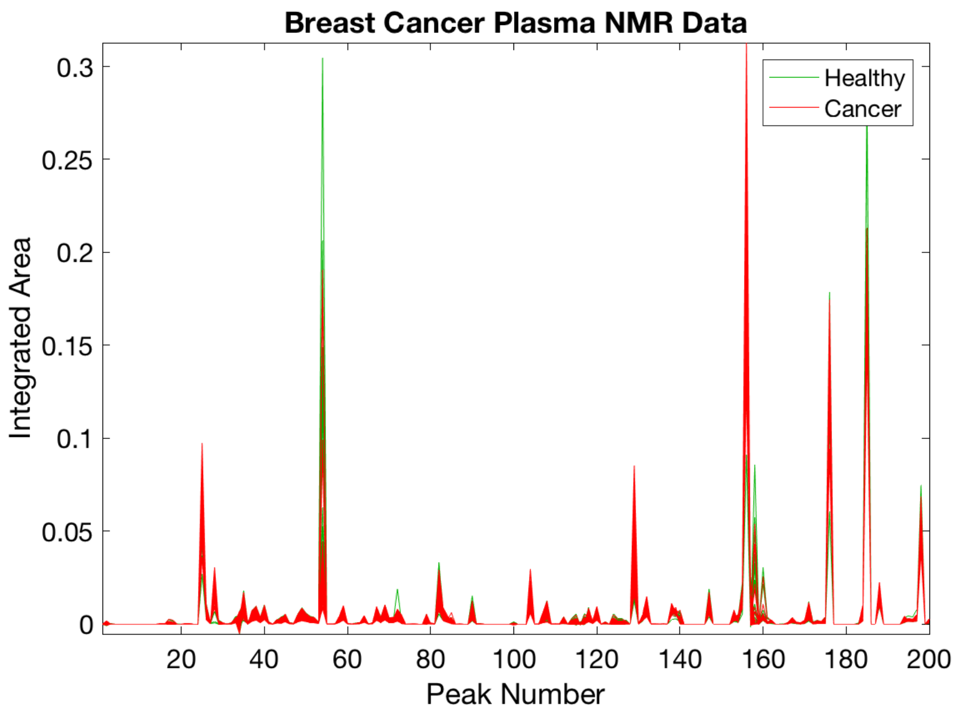


- 1-within normal limits
- 2-normal squamous
- 3-normal columnar
- 4-squamous metaplasia
- 5-LoSIL
- 6-HiSIL

	Calibration	Test
Normal Squamous	152	33
Squamous Metaplasia (pre-cancerous) LoSIL (low-grade cancerous) HiSIL (high-grade cancerous)	244	66

EIGENVECTOR RESEARCH INCORPORATED

Breast Cancer Forecasting



define as a complex pattern of relevant biological and

- 883 Danish women, half diagnosed with breast cancer
- Plasma samples taken years before diagnosis at beginning of study, then stored
- Analyzed by proton NMR, peaks integrated

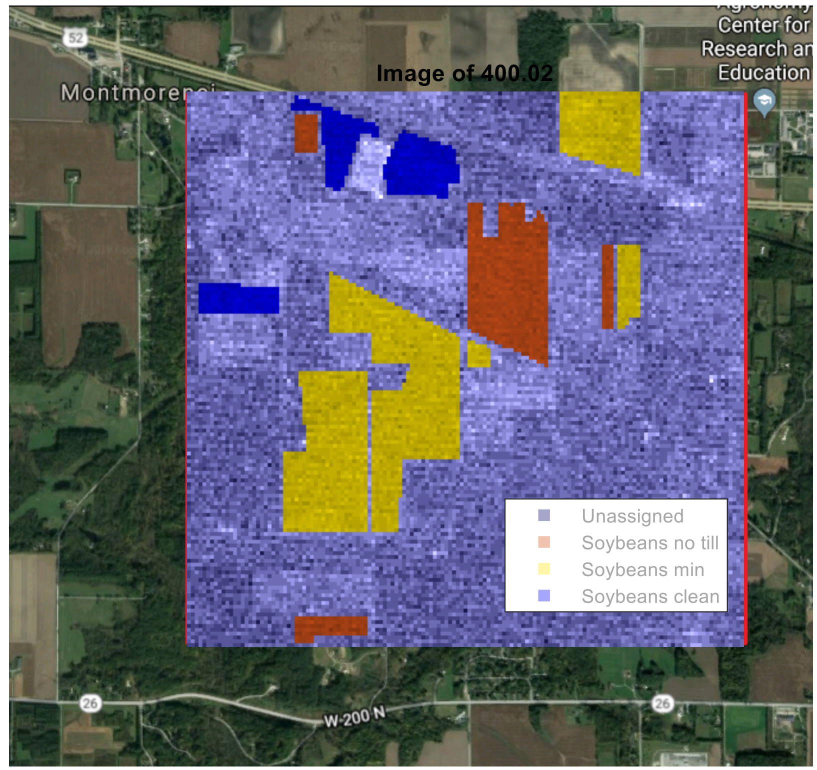
EIGENVECTOR RESEARCH INCORPORATED

Infectious Disease Detection

- Bacteria separated
- Measured with Excitation Emission Fluorescense
- Unfolded, 670 variables
- Goal is to predict if bacteria level is above a threshold value
- 1155 Calibration samples, 58% positive
- 385 Test samples, 60% positive

The IndianPines Dataset

- Hyperspectal image of a mixed farmland area west of Lafayette, Indiana.
- 145x145 pixels
- 220 spectral channels
- Use only pixels from the Soy fields, which are of 3 types: “No till”, “Min” and “Clean”.
- (“Min” = “Min till”)



EIGENVECTOR RESEARCH INCORPORATED

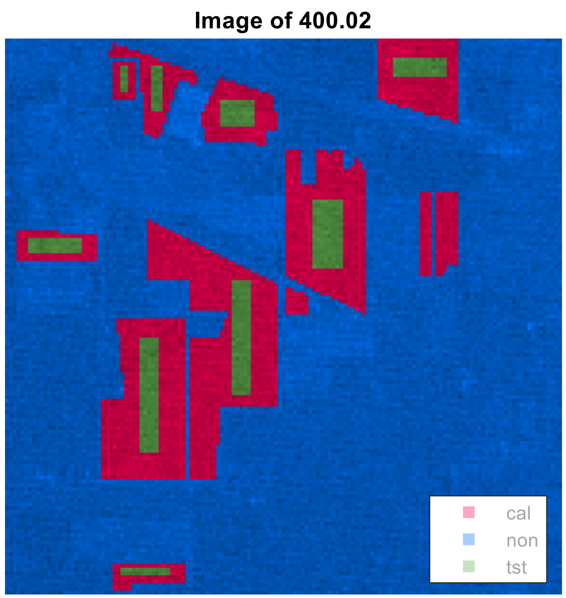
Data Details

Soy fields types: “No till”, “Till”, “Clean”

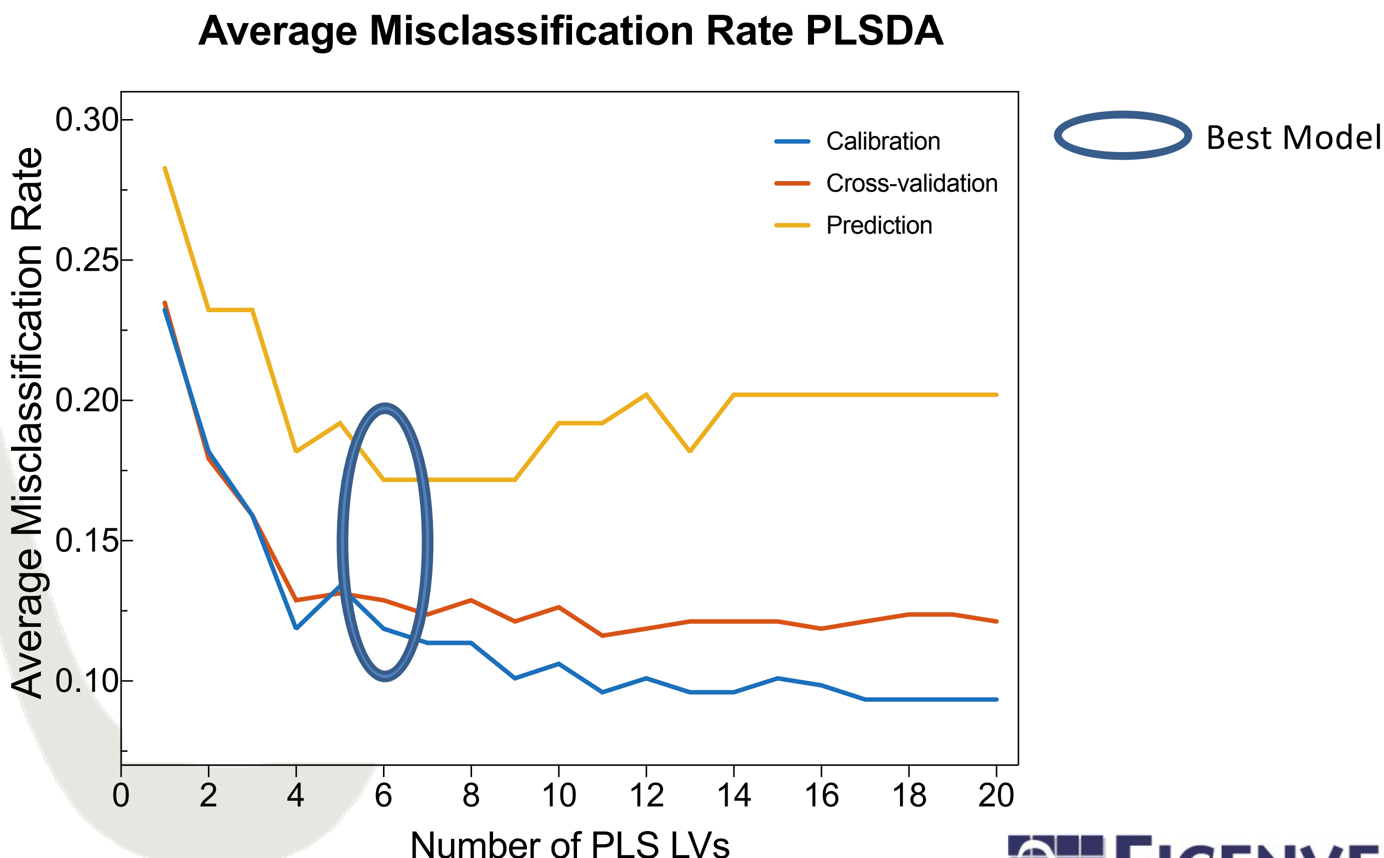
4050 Soy field pixels used

82% as Calibration, 18% as Test where test pixels are contiguous areas within a Soy field

```
No Till: 24%  (968 pixels, 784 cal, 184 test)
Min:     61% (2468 pixels; 2098 cal, 370 test)
Clean:   15%  (614 pixels; 459 cal, 155 test)
```



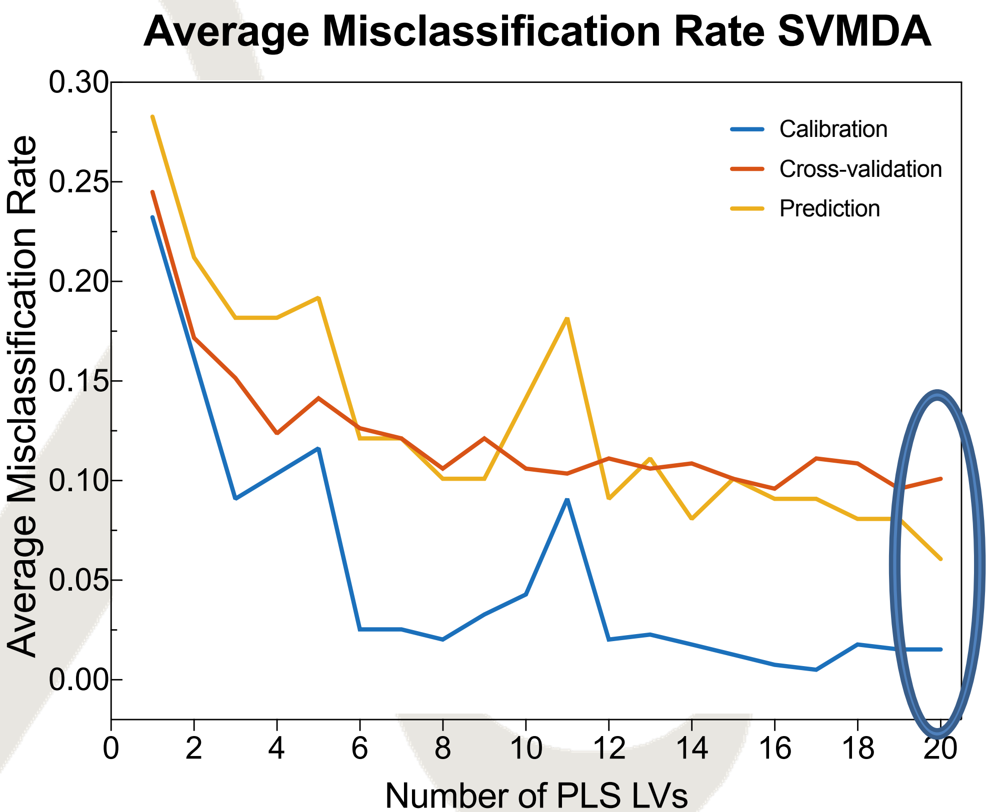
PLS-DA – Cervical Cancer



Autoscaled, 5-fold Cross-validation

EIGENVECTOR RESEARCH INCORPORATED

PLS/SVM-DA – Cervical Cancer



Autoscaled, 5-fold Cross-validation

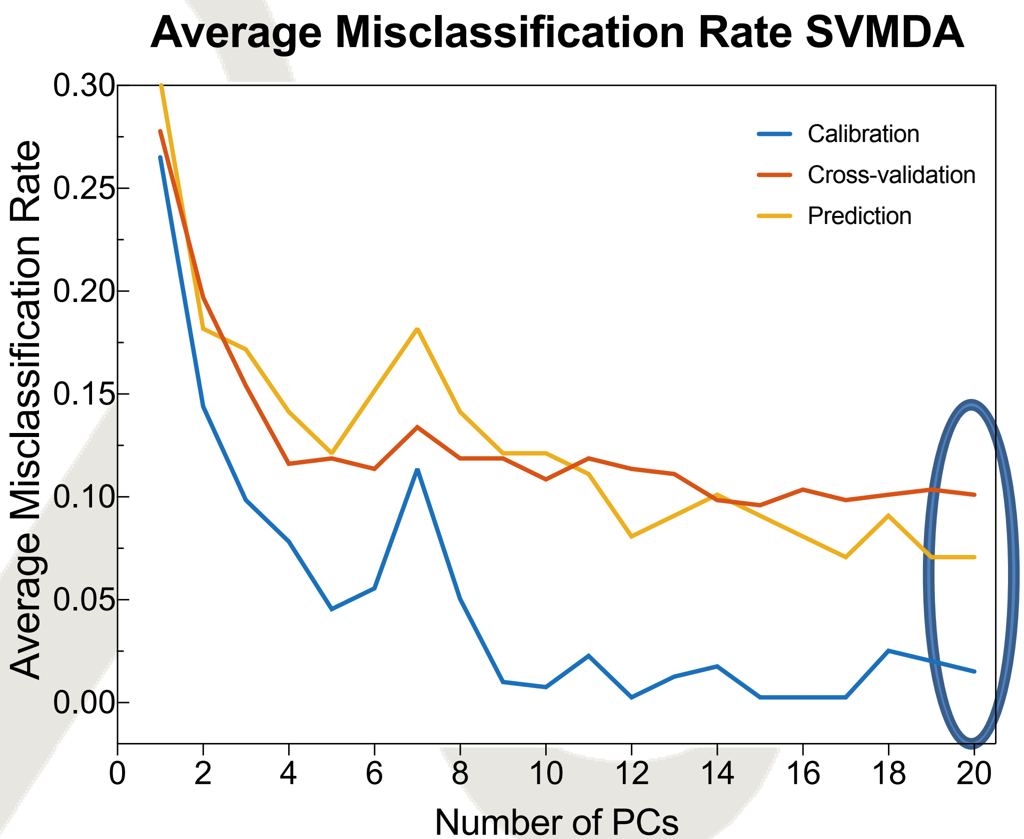
Best Model

Average Misclassification without compression

Calibration Error	= 0.08
Cross-validation Error	= 0.11
Prediction Error	= 0.16

EIGENVECTOR RESEARCH INCORPORATED

PCA/SVM-DA – Cervical Cancer



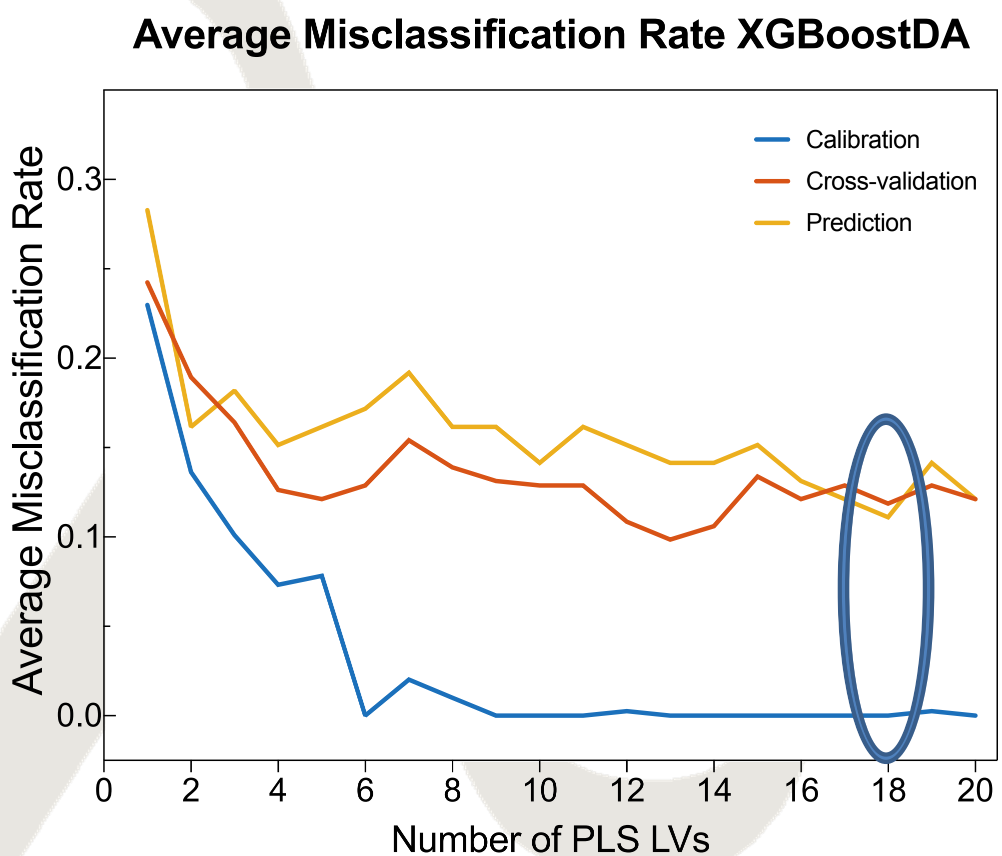
Autoscaled, 5-fold Cross-validation

Best Model

Average Misclassification without compression
Calibration Error = 0.08
Cross-validation Error = 0.11
Prediction Error = 0.16

EIGENVECTOR RESEARCH INCORPORATED

PLS/XGBoostDA – Cervical Cancer



Autoscaled, 5-fold Cross-validation

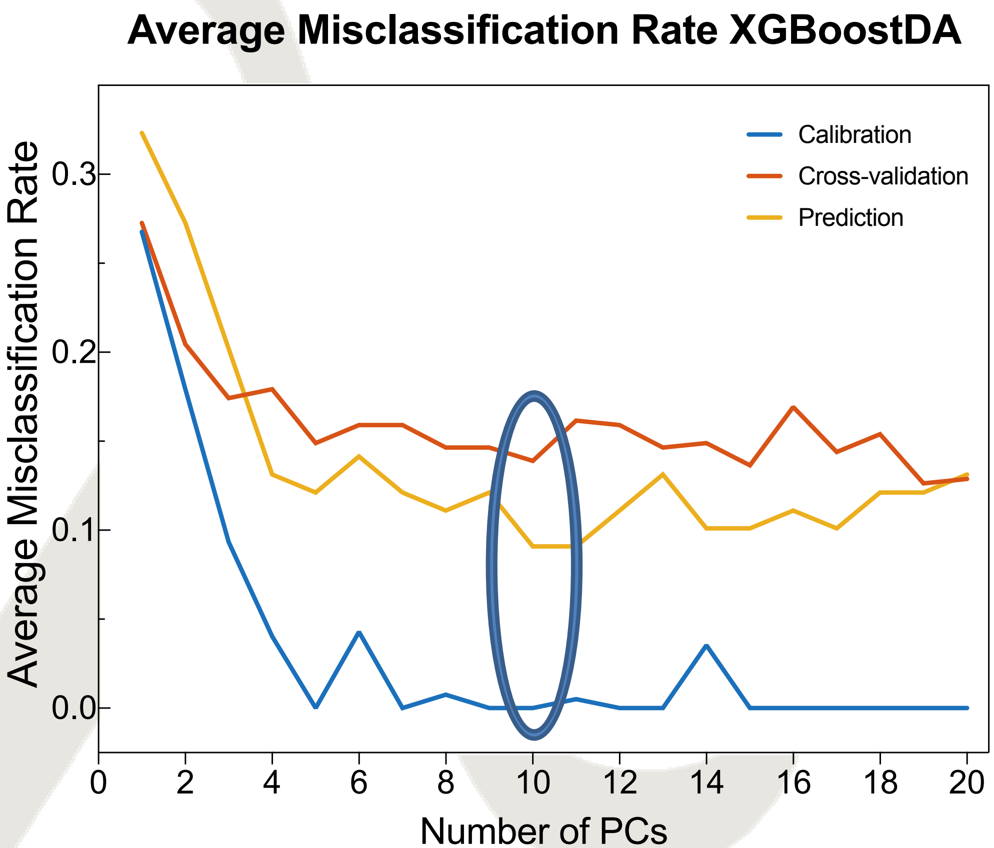
Best Model

Average Misclassification without compression

Calibration Error	= 0.00
Cross-validation Error	= 0.14
Prediction Error	= 0.11

EIGENVECTOR RESEARCH INCORPORATED

PCA/XGB-DA – Cervical Cancer



Autoscaled, 5-fold Cross-validation

Best Model

Average Misclassification without compression
Calibration Error = 0.00
Cross-validation Error = 0.14
Prediction Error = 0.11

EIGENVECTOR RESEARCH INCORPORATED

Cervical Cancer Summary

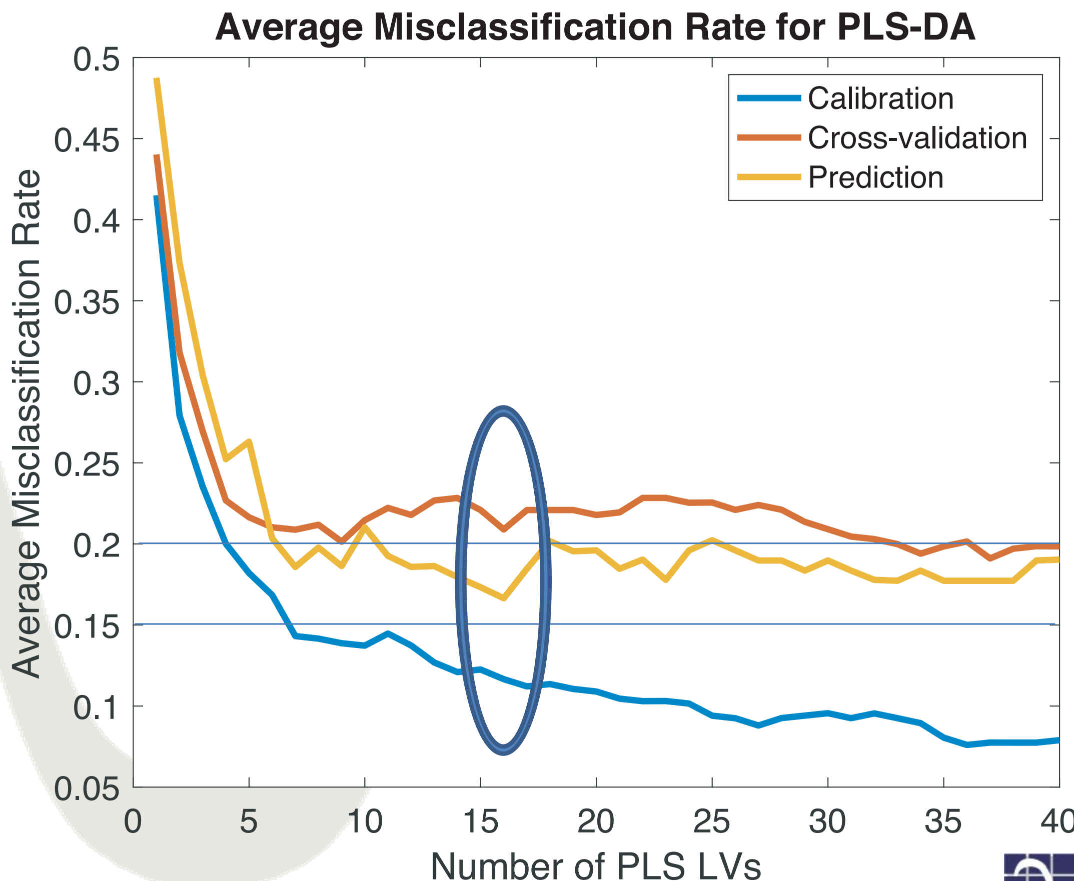
	Best models w/PLS		Best models w/PCA		No Compress
	Numer of LVs	Error	Numer of LVs	Error	Error
PLSDA	6	0.17	-	-	-
SVMDA*	20	0.06	17	0.07	0.16
XGBoostDA	18	0.11	10	0.09	0.11

*The total number of variables is 22

- PLSDA < XGBoostDA < SVMDA
- SVMDA Performs much better with compression at almost full rank, but also better in the compressed subspace.
- XGBoostDA seems less sensitive to compression.
- XGBoostDA almost always overfits the calibration, but cross-validation consistently shows a good estimation of the actual performance of the models when compared to the test set.

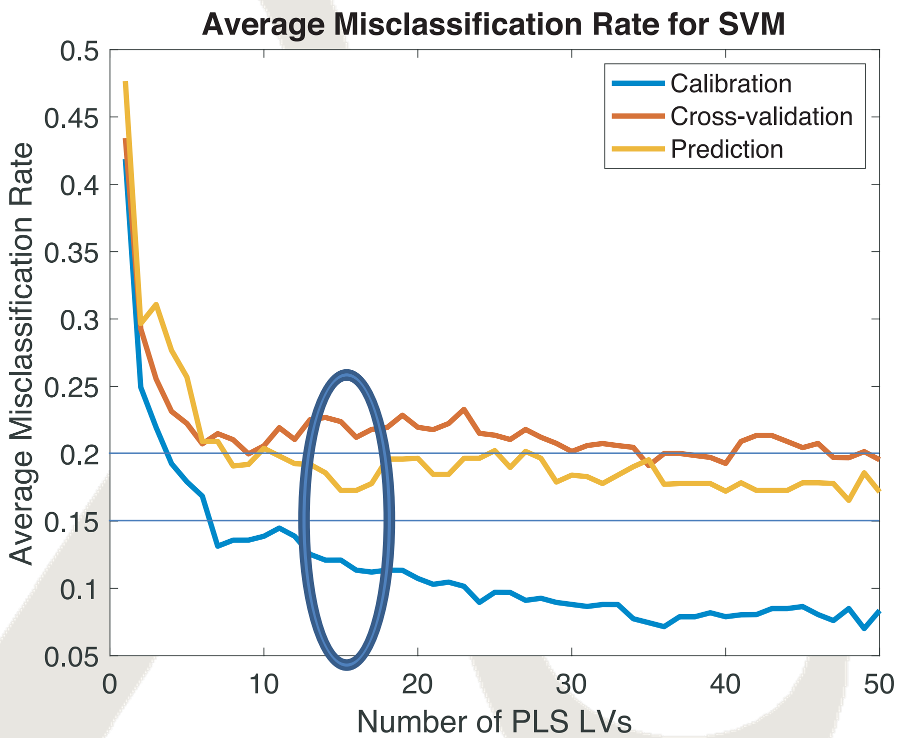
EIGENVECTOR RESEARCH INCORPORATED

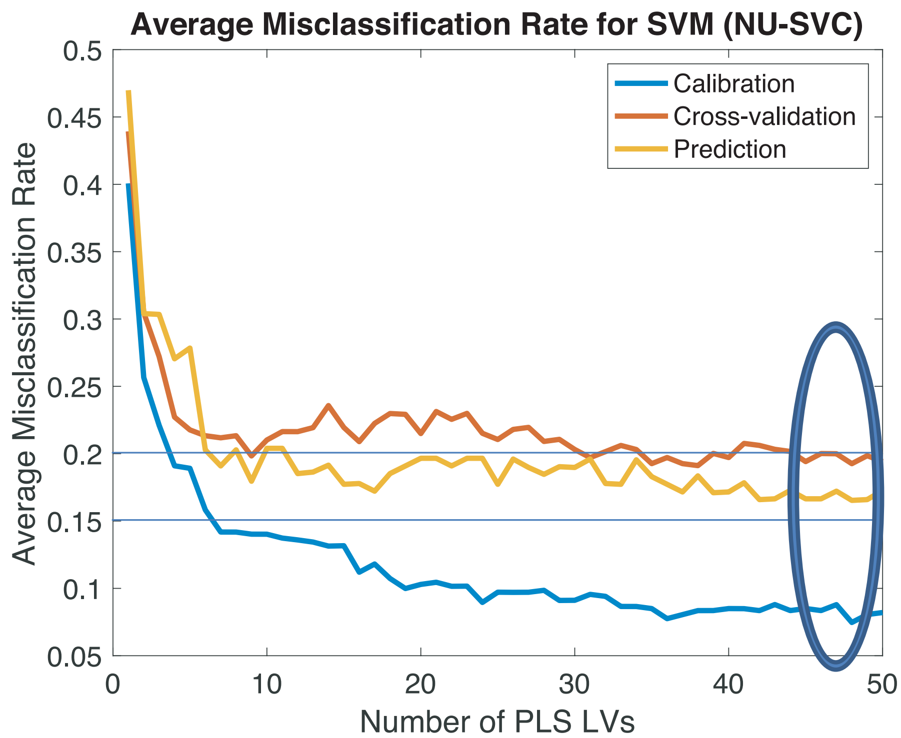
Breast Cancer Detection Results



EIGENVECTOR RESEARCH INCORPORATED

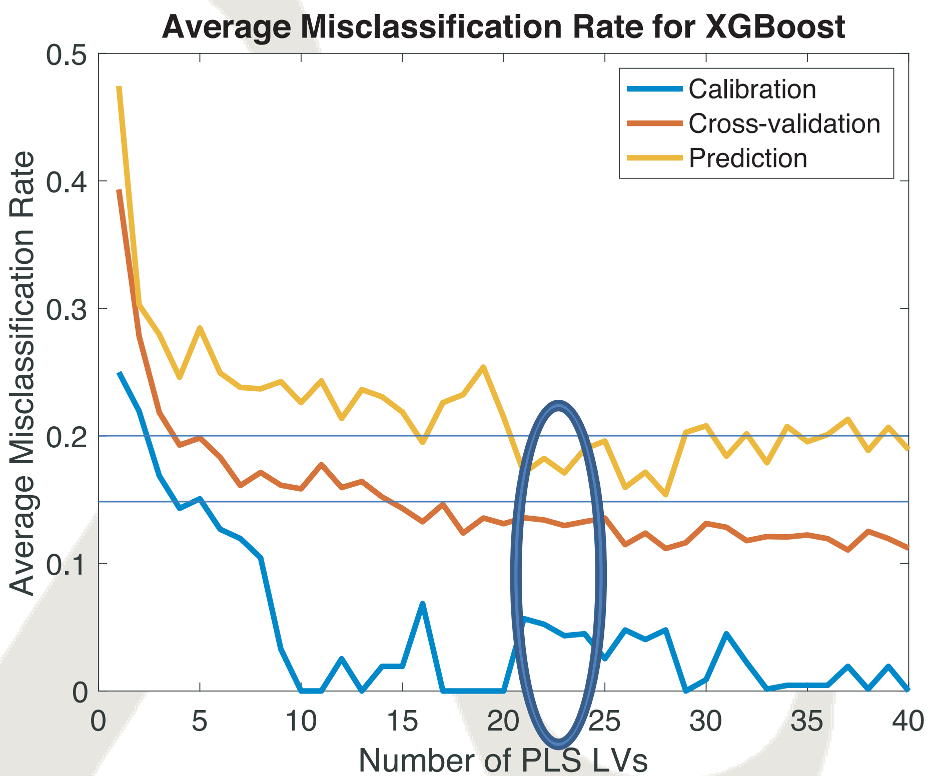
SVM-DA on Breast Cancer

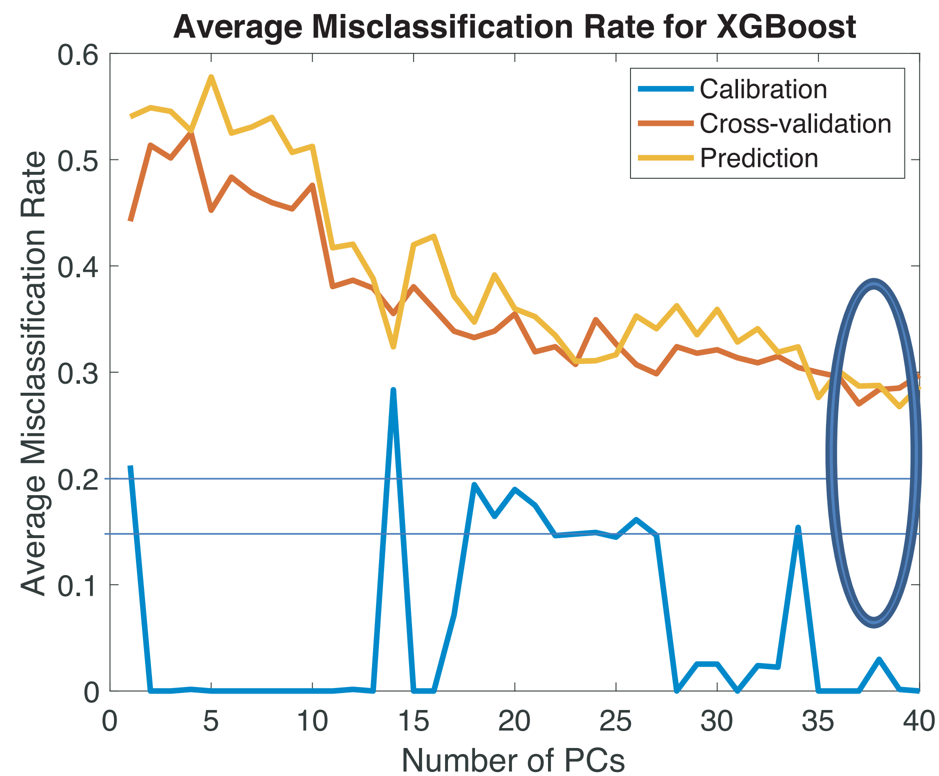




EIGENVECTOR RESEARCH INCORPORATED

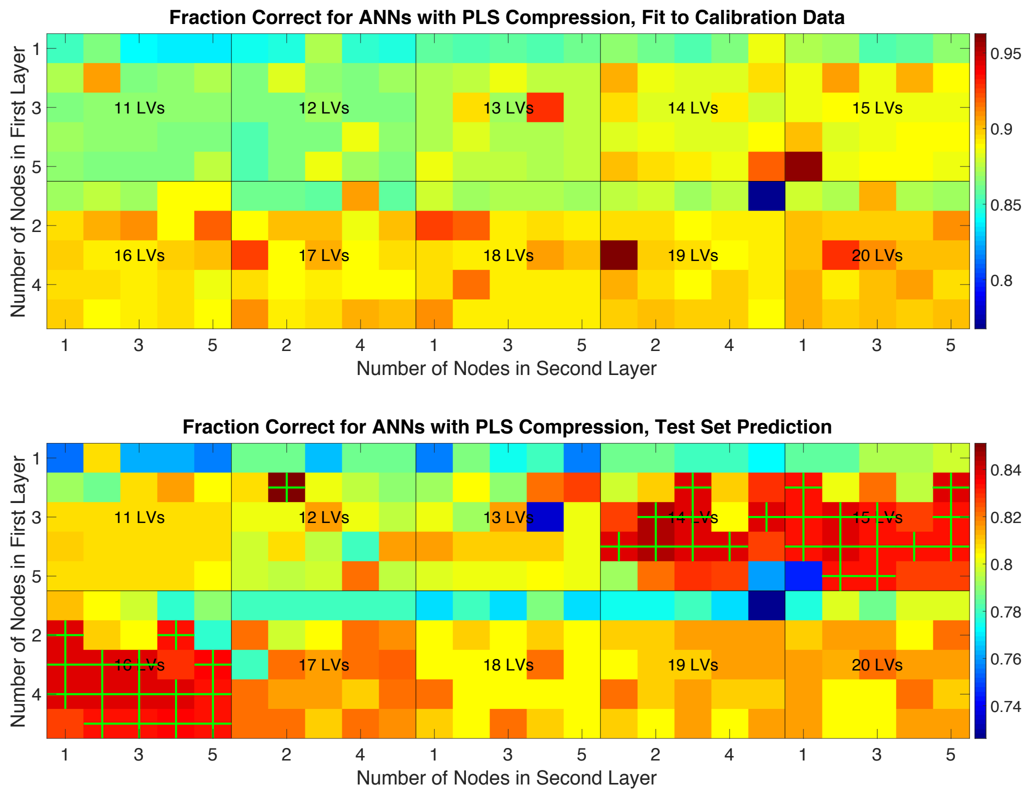
XGB-DA on Breast Cancer





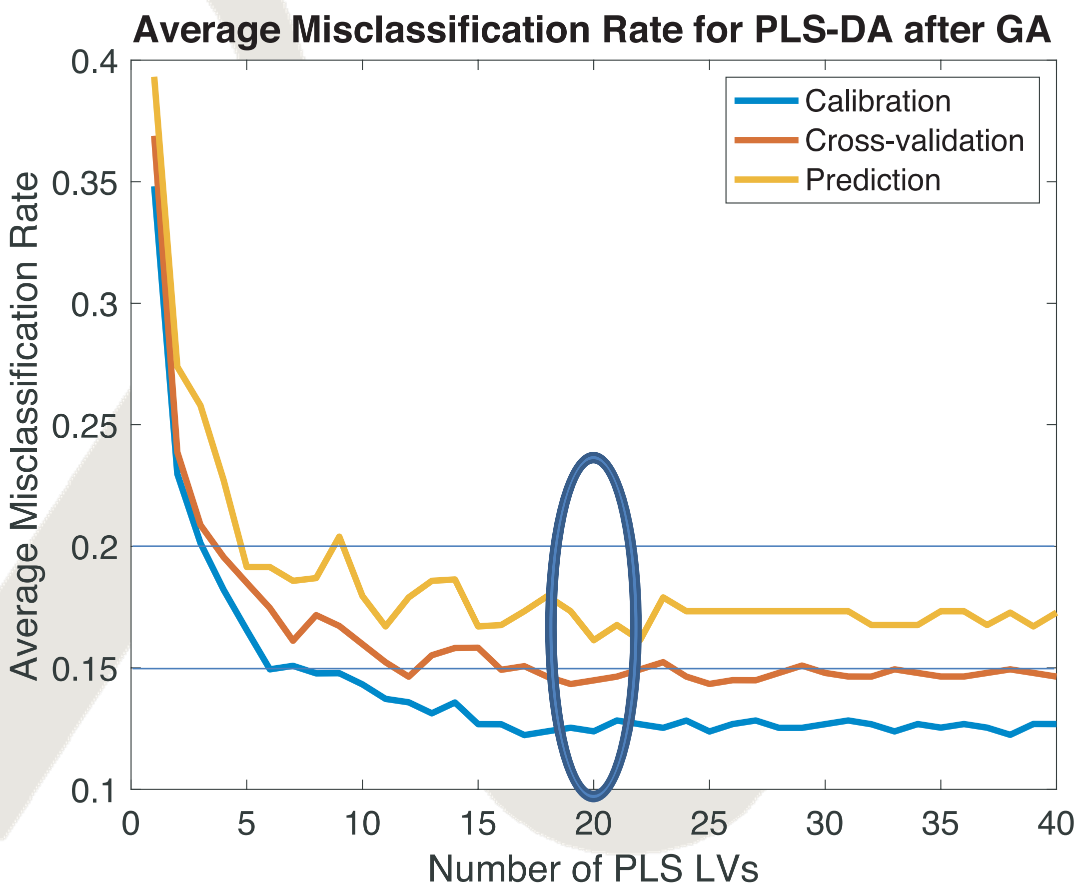
EIGENVECTOR RESEARCH INCORPORATED

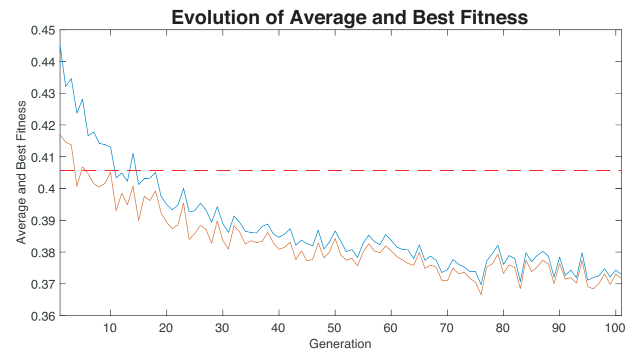
ANN on Breast Cancer

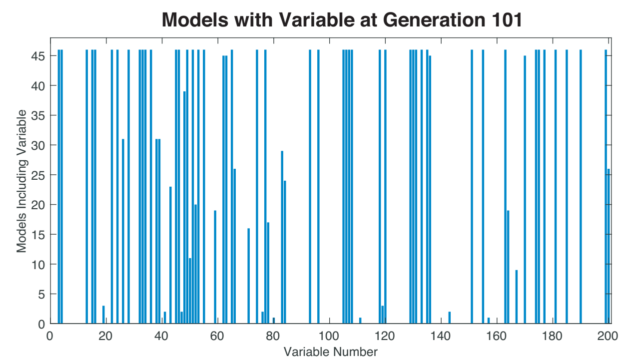


EIGENVECTOR RESEARCH INCORPORATED

PLS-DA GA on Breast Cancer







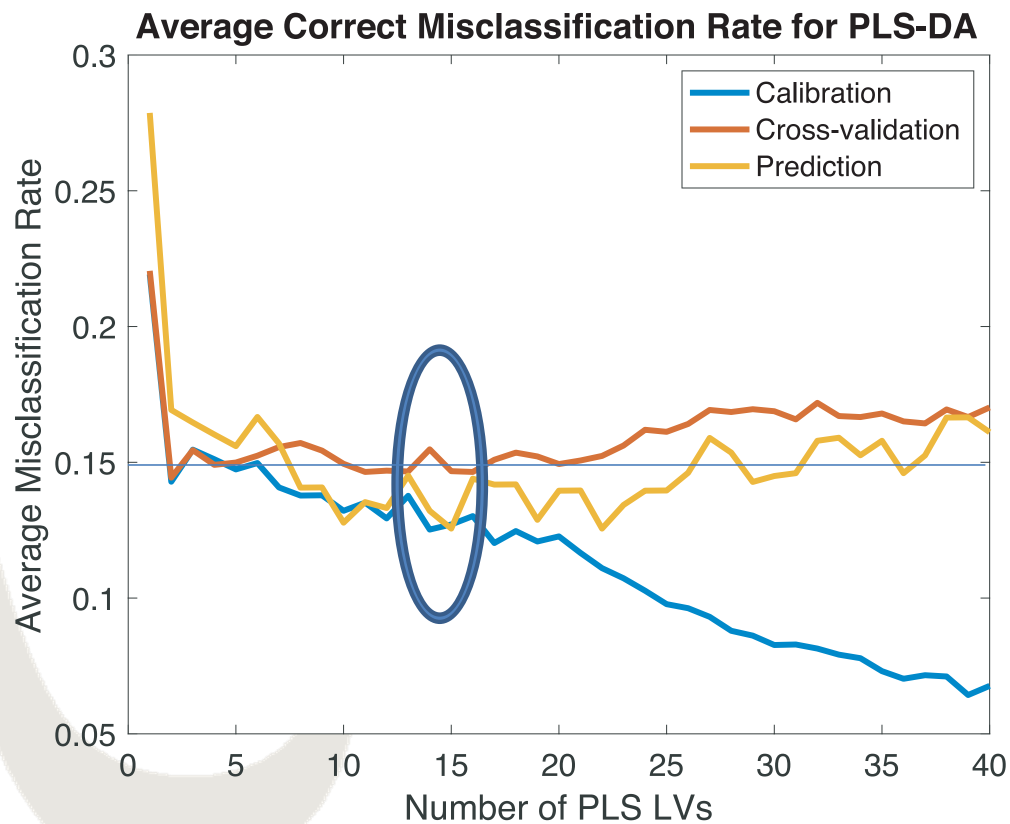
EIGENVECTOR RESEARCH INCORPORATED

Summary of Breast Cancer Resuts

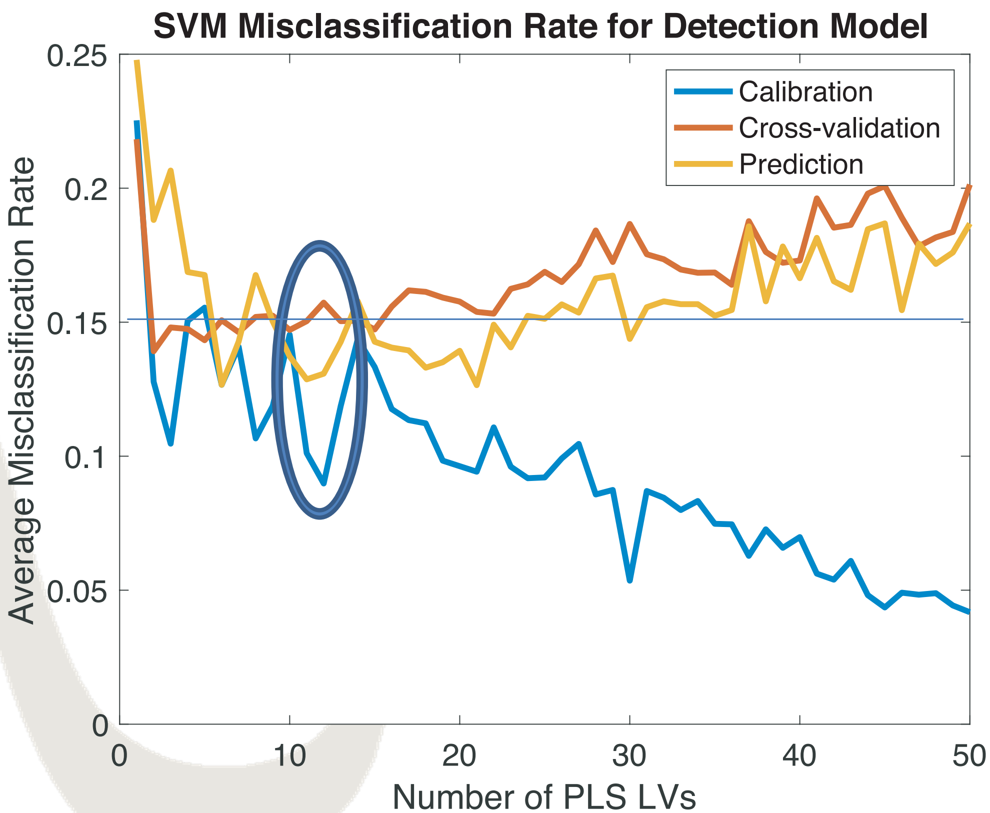
- Compression is important in ANN, SVM, XGBoost
- All methods able to achieve error of ~0.17
- Success of each depends on final criteria for model selection
  - Which model do you choose?
- ANN had most models around best performance
- PLS-DA with variable selection strong contender

EIGENVECTOR RESEARCH INCORPORATED

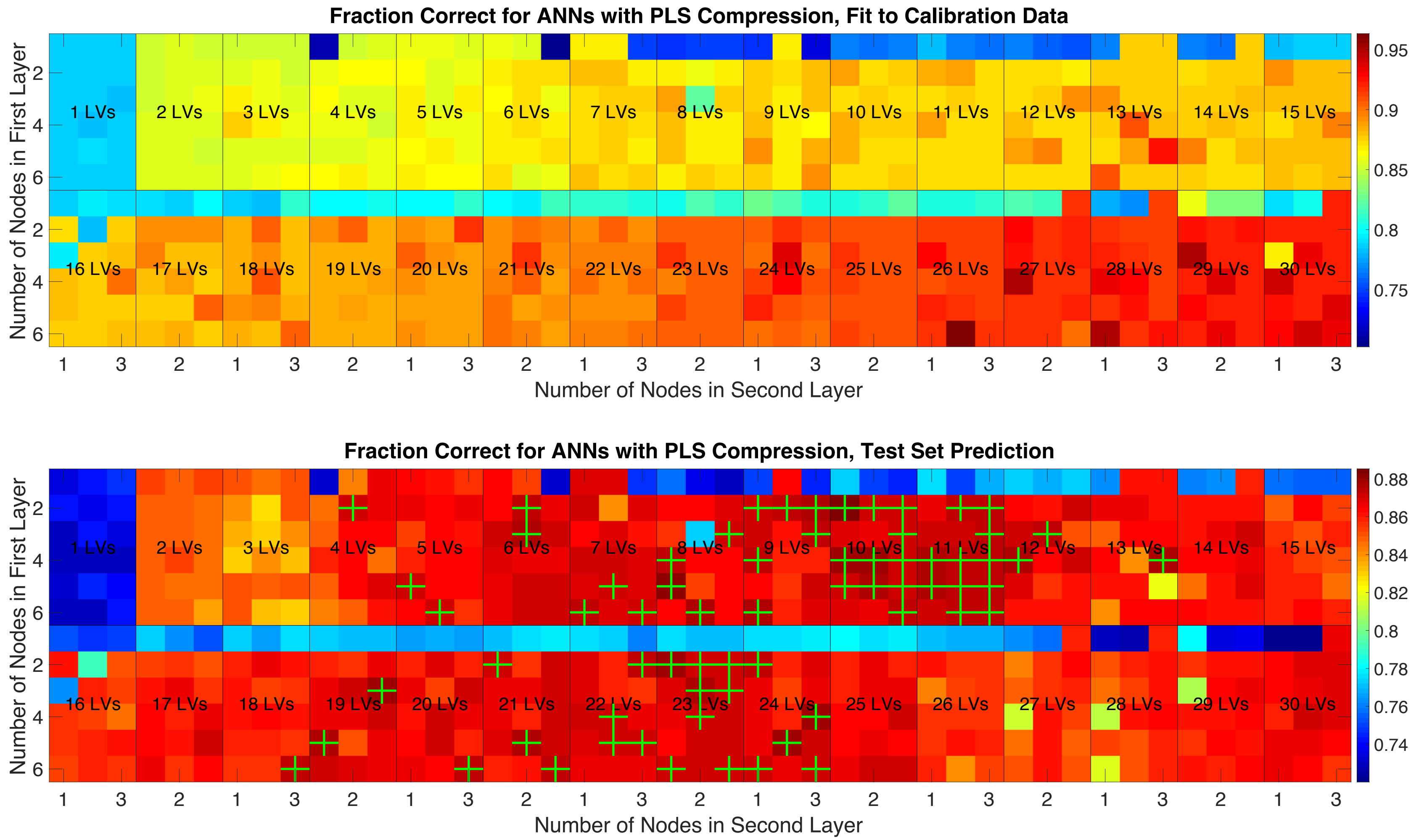
Disease Detection Results PLS-DA



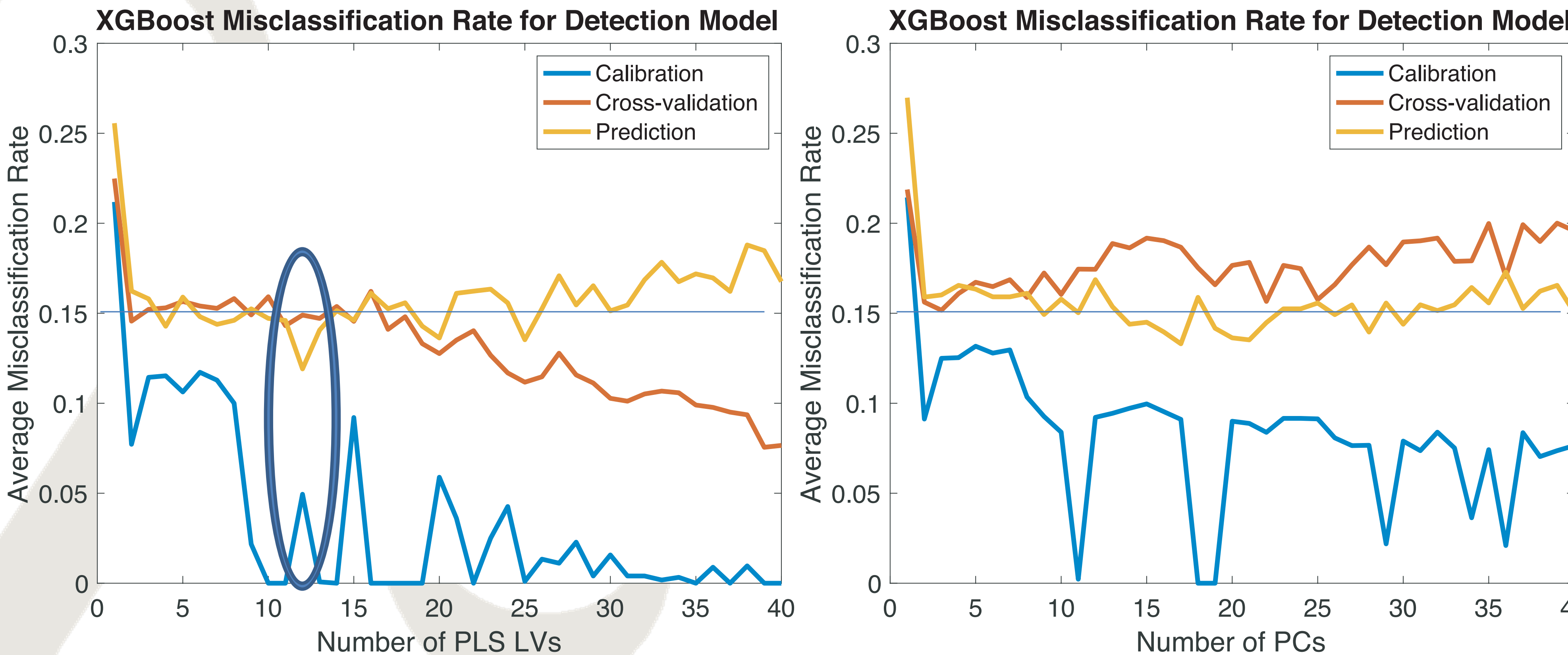
SVM-DA on Disease Detection



EIGENVECTOR RESEARCH INCORPORATED


Fraction Correct for ANNs with PLS Compression, Fit to Calibration Data
Number of Nodes in First Layer
Number of Nodes in Second Layer
1 LVs
2 LVs
3 LVs
4 LVs
5 LVs
6 LVs
7 LVs
8 LVs
9 LVs
10 LVs
11 LVs
12 LVs
13 LVs
14 LVs
15 LVs
16 LVs
17 LVs
18 LVs
19 LVs
20 LVs
21 LVs
22 LVs
23 LVs
24 LVs
25 LVs
26 LVs
27 LVs
28 LVs
29 LVs
30 LVs
0.95
0.9
0.85
0.8
0.75
Fraction Correct for ANNs with PLS Compression, Test Set Prediction
Number of Nodes in First Layer
Number of Nodes in Second Layer
0.88
0.86
0.84
0.82
0.8
0.78
0.76
0.74

XGB-DA on Disease Detection

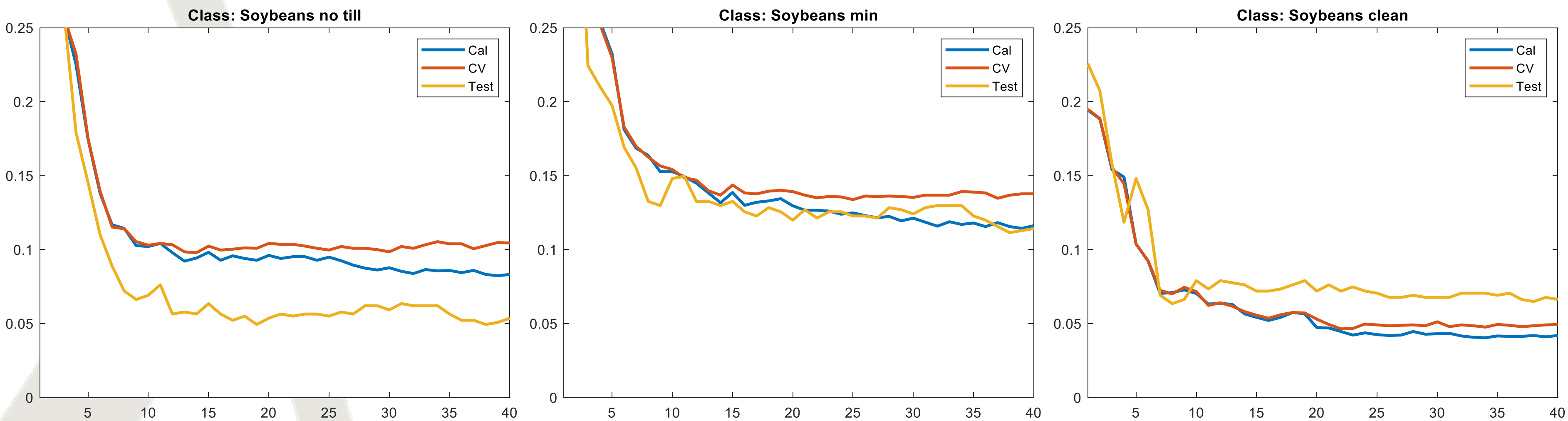


EIGENVECTOR RESEARCH INCORPORATED

Disease Detection Results Summary

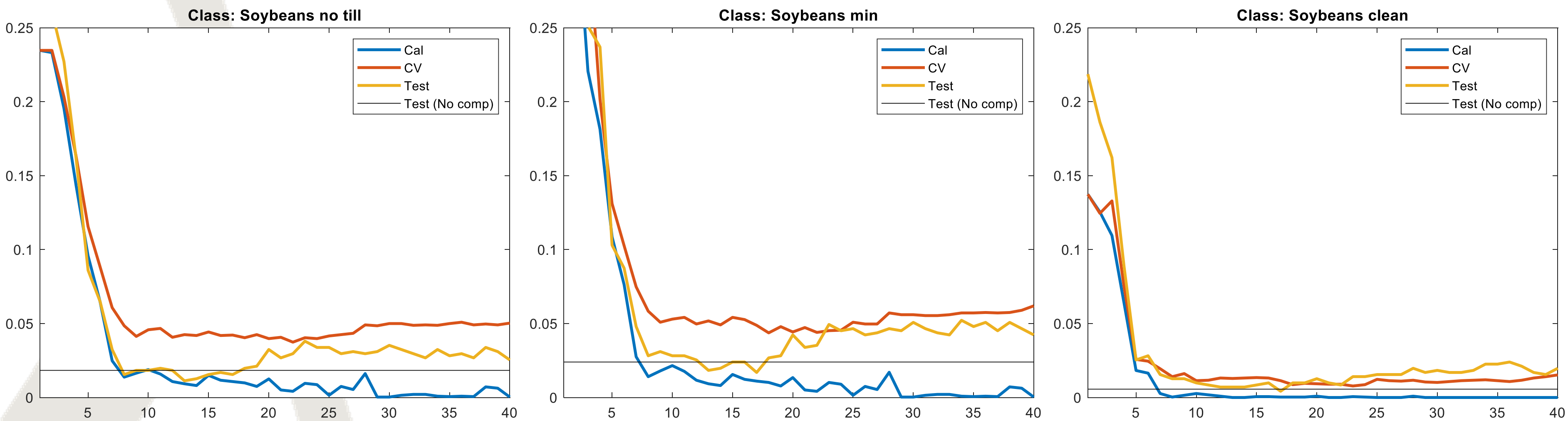
- All methods benefited from compression
  - PLS compression worked better than PCA
- Best error rate ~0.13 for all methods

PLS-DA on Crop Identification



Compression LVs: 1:40
{'derivative', 'snv', 'mean center’}

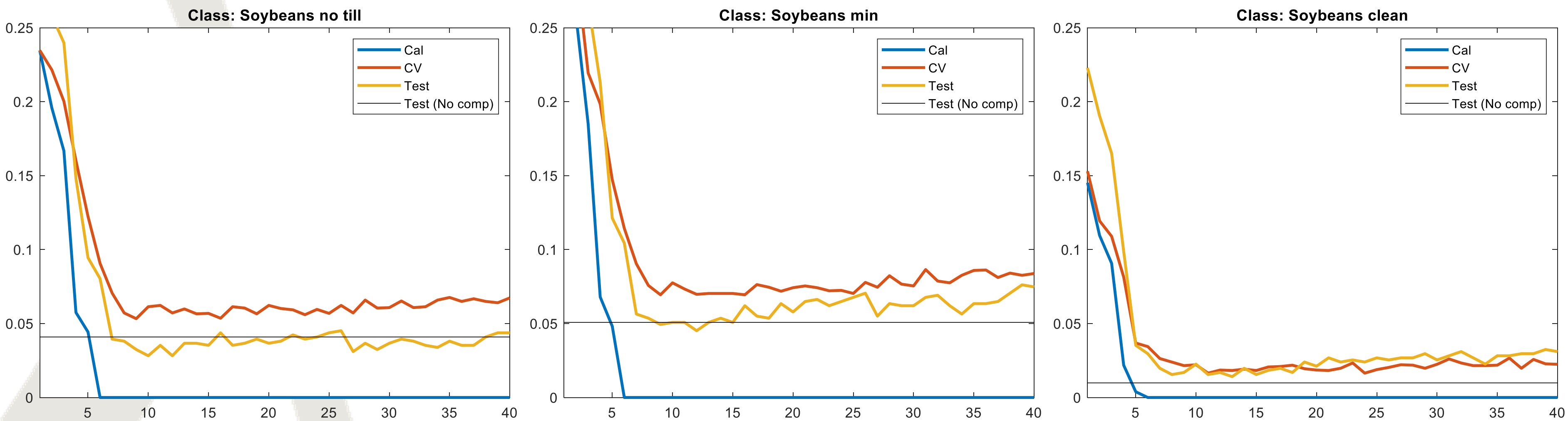
SVM-DA on Crop Identification



Compression LVs: 1:40
Optimize over full parameter range
{'derivative', 'snv', 'mean center'});

```
No compression (Black horiz. line shows Test Error):
                                No Fill      Min    Clean
misclassification (CV):    0.0422, 0.0500, 0.0138
misclassification (Test): 0.0183, 0.0240, 0.0056
```

XGB-DA on Crop Identification



Compression LVs: 1:40
Optimize over full parameter range
{'derivative', 'snv', 'mean center'});

```
No compression (Black horiz. line shows Test error):
                                 No Fill    Min    Clean
misclassification (CV):    0.0425, 0.0551, 0.0174
misclassification (Test): 0.0409, 0.0508, 0.0099
```

EIGENVECTOR RESEARCH INCORPORATED

Crop Identification Summary

- SVMDA gives the best performance on the Validation data for all 3 classes.
- SVMDA & XGBDA much better than PLSDA for CV or Validation data
- SVMDA and XGBDA behave similarly when PLS compression is used where error decreases rapidly up to about 10 LVs used, approximately matching no compression, then deteriorating when >15 LVs used in compression

Practical Considerations

- PLS-DA much faster than other methods
  - Allows exploration of wider preprocessing space
  - Has better diagnostics, more interpretable
- Compression speeds up other methods considerably

Overall Summary

- Useful to explore parameter options, especially compression
- SVM-DA overall winner
  - But didn't do ANNs on cervical cancer and crop detection
- XGB-DA always overfits calibration data
  - But cross-validation results largely agree with prediction results

Often, the problem is the data!

EIGENVECTOR RESEARCH INCORPORATED