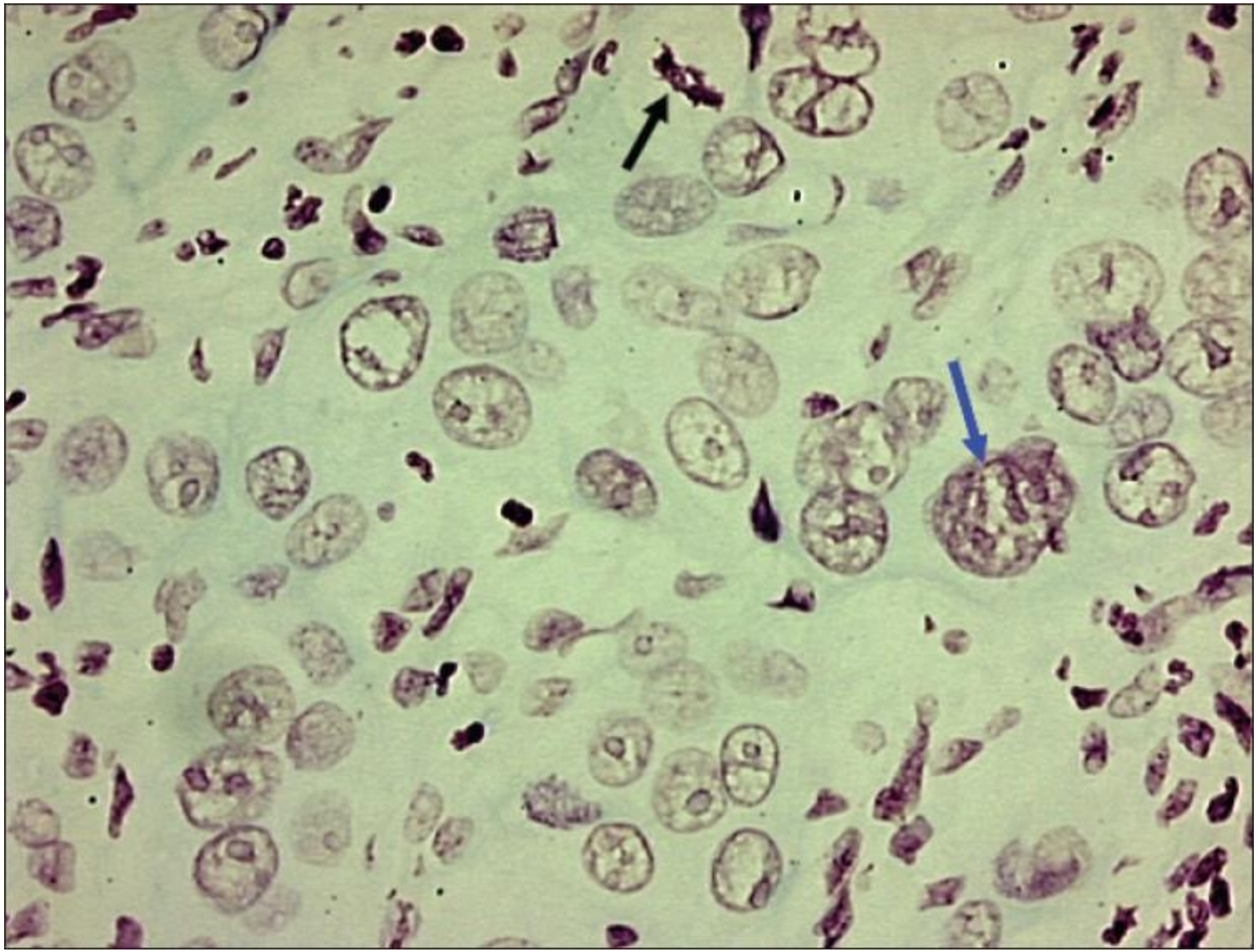
Mixture of Gaussian distribution

Simulating D.I. values

Objectives

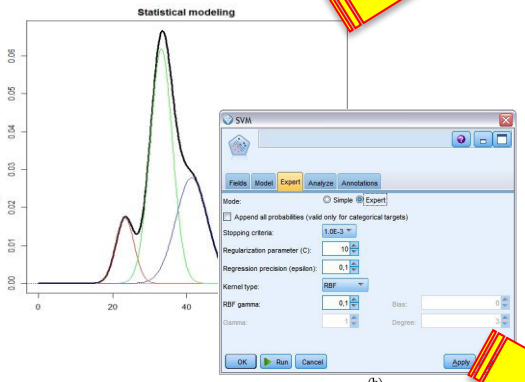
- Understanding and unveiling the biological mechanism of oral squamous cell carcinoma
- Leveraging nuclei stain imaging technology (Feulgen staining)
- Extracting the determining digitized information
- Developing data processing protocol
- Building predicting model for detecting abnormal cell dividing – cancer diagnosis
- Establishing the standard for oral cancer early diagnosis

Photomicrograph of moderately well differentiated OSCC showing large tumor nucleus with multiple prominent nucleoli (blue arrow) and abnormal mitotic figure (black arrow)



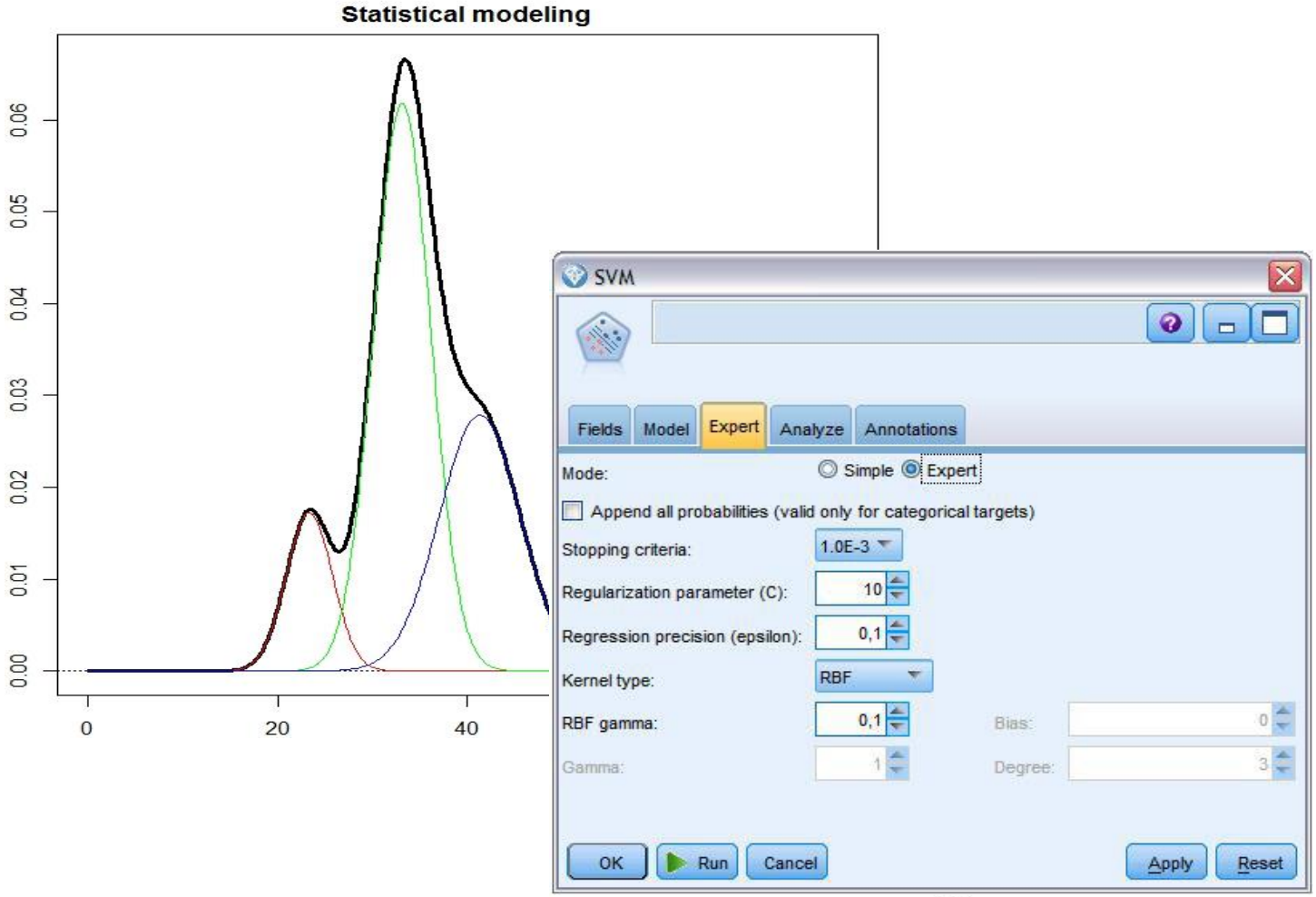
DB Nandini and RV Subramanyam, 2011, Nuclear features in oral squamous cell carcinoma: A computer-assisted microscopic study, V.15:2, 177-181,JOMFP

Raw dataDB


Statistical modeling
SVM
Fields
Model
Expert
Analyze
Annotations
Mode:
Simple
Expert
Append all probabilities (valid only for categorical targets)
Stopping criteria:
Regularization parameter (C):
Regression precision (epsilon):
Kernel type:
RBF gamma:
Bias:
Degree:
OK
Run
Cancel
Apply

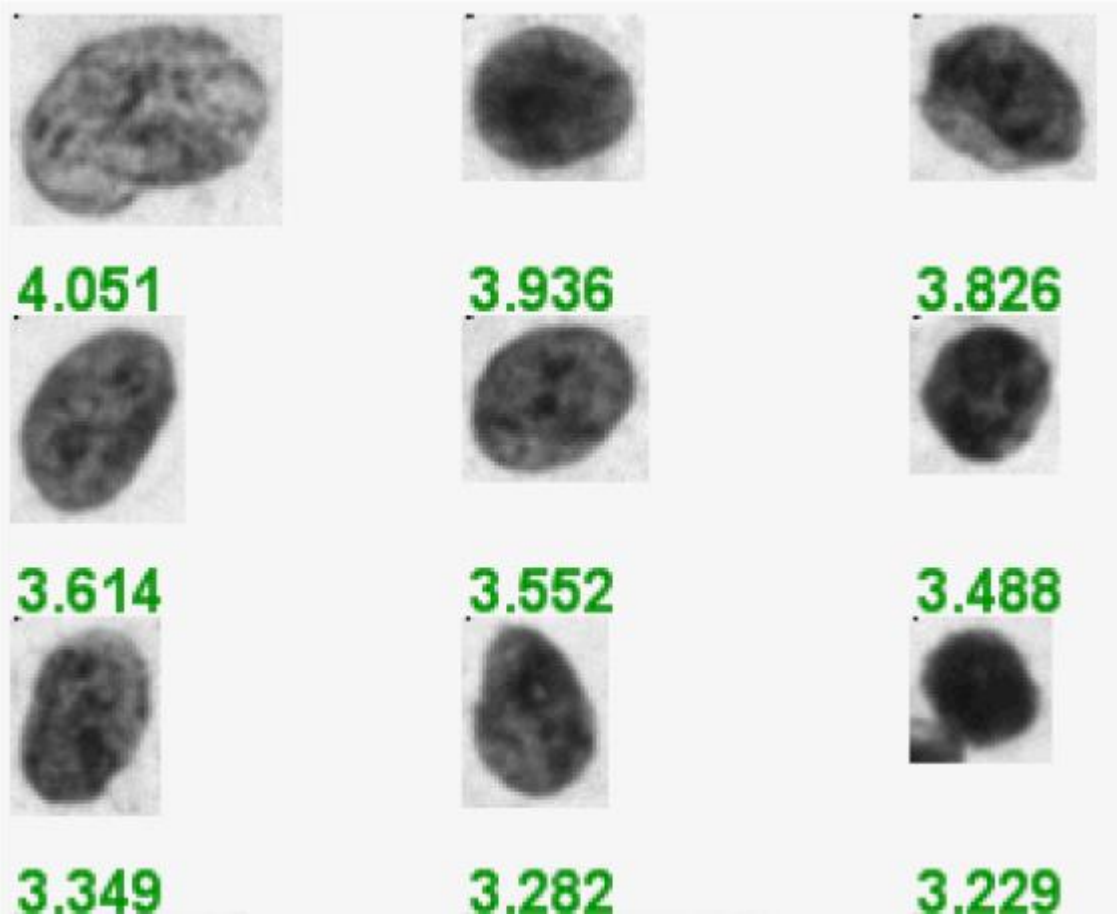
(b)


dreamstime


Statistical modeling
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0
20
40
SVM
Fields
Model
Expert
Analyze
Annotations
Mode:
Simple
Expert
Append all probabilities (valid only for categorical targets)
Stopping criteria:
1.0E-3
Regularization parameter (C):
10
Regression precision (epsilon):
0,1
Kernel type:
RBF
RBF gamma:
0,1
Bias:
0
Gamma:
1
Degree:
3
OK
Run
Cancel
Apply
Reset

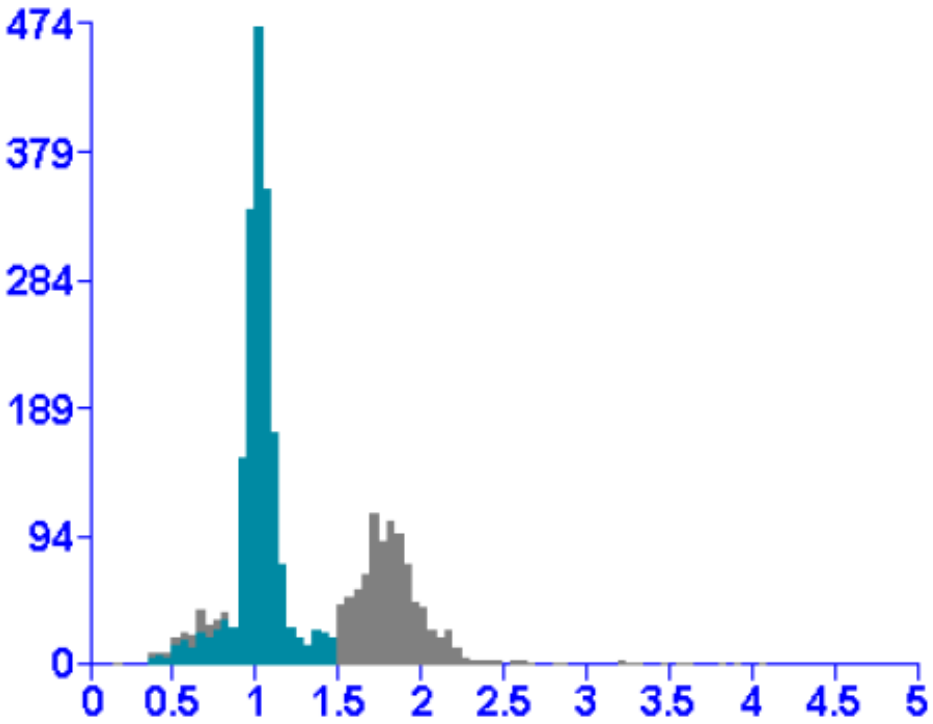
(b)

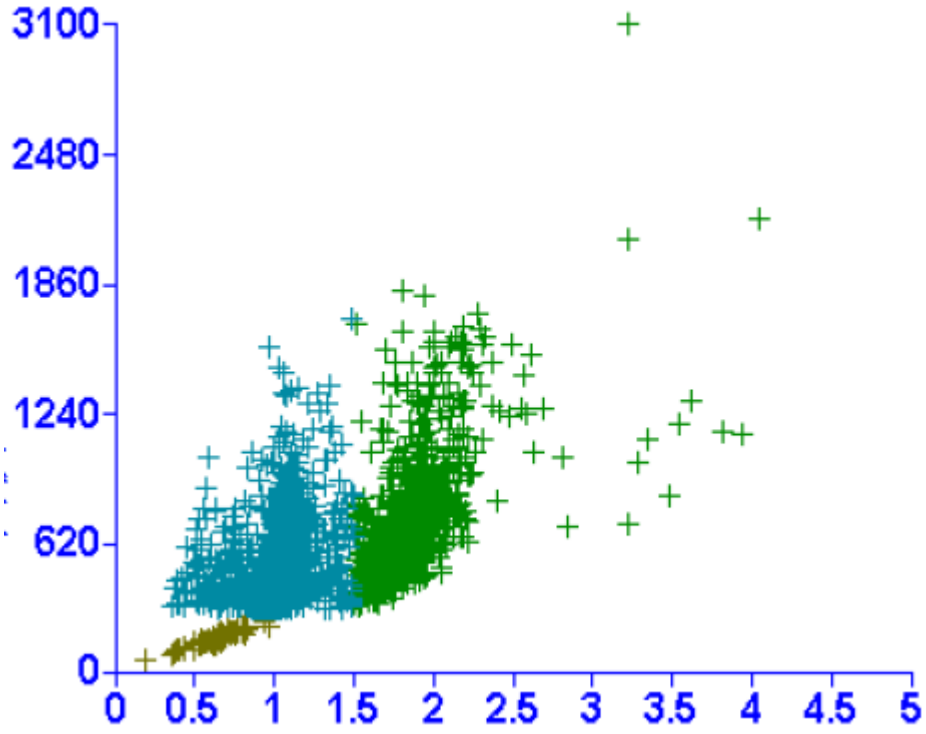
Abnormal nucleus image (D.I. value)



Papanicolaou (Pap) smears and nucleus was stained with Feulgan stain

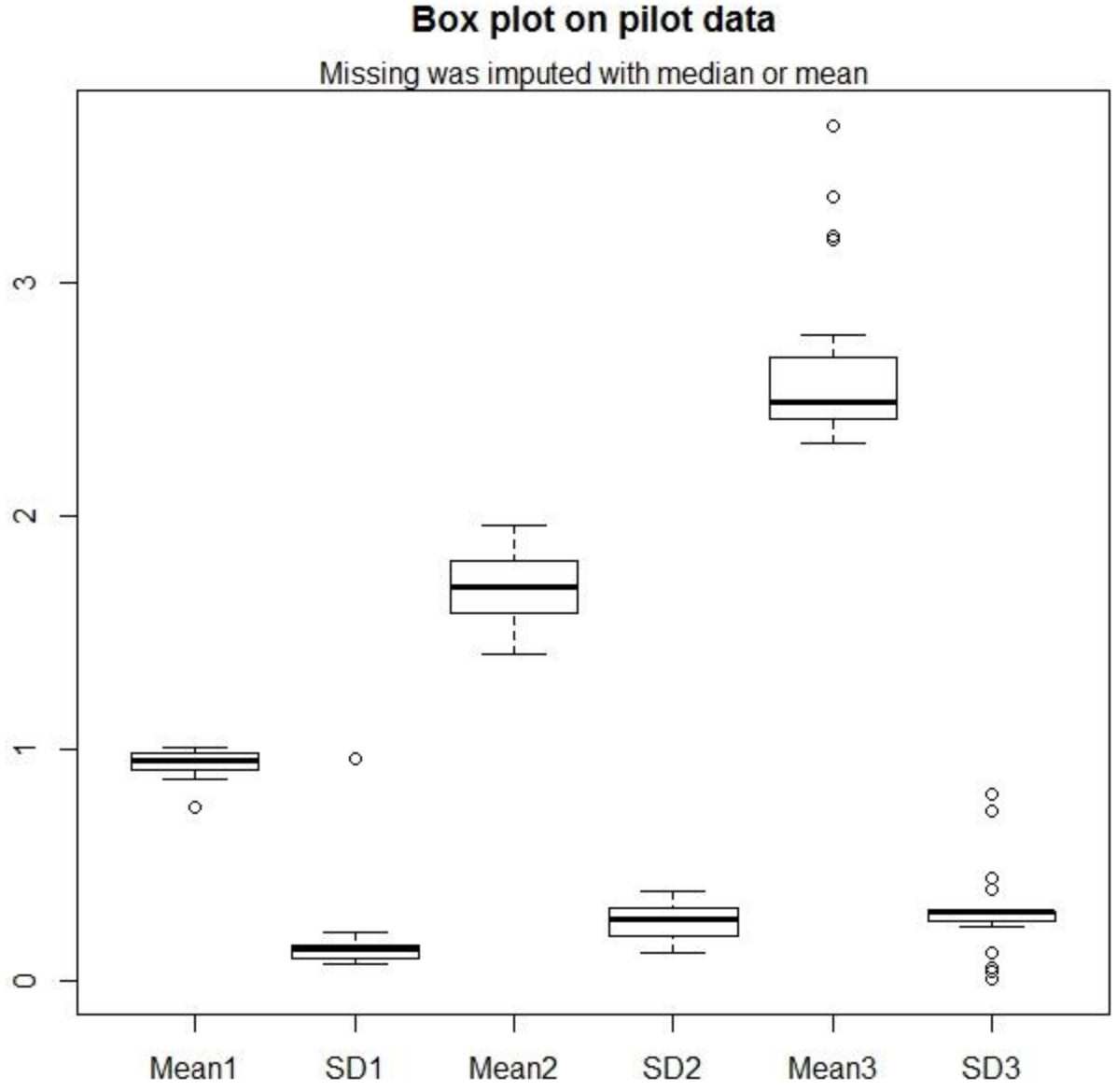
Cell population distribution





Summary statistics

Group	Mean1	SD1	Mean2	SD2	Mean3	SD3
normal1	0.96	0.155	1.961	0.22	2.321	NA
normal2	0.878	0.178	1.675	0.361	2.355	NA
normal3	0.925	0.153	1.642	0.349	NA	NA
normal4	1.001	0.088	1.878	0.237	NA	NA
normal5	0.96	0.133	1.778	0.267	2.307	0.009
normal6	0.974	0.1	1.588	0.32	NA	NA
normal7	0.888	0.182	1.407	0.278	2.391	0.046
normal8	0.963	0.155	1.71	0.256	NA	NA
normal9	0.995	0.085	1.434	0.294	NA	NA
OLK1	0.751	0.137	1.578	0.172	3.675	NA
OLK2	0.959	0.117	1.718	0.368	NA	NA
OLK3	0.94	0.14	1.938	0.19	NA	NA
OLK4	0.992	0.074	1.66	0.172	3.2	0.261
OLK5	0.912	0.155	1.818	0.272	3.366	0.808
OLK6	0.949	0.138	1.798	0.315	2.332	NA
OLK7	0.95	0.129	1.771	0.27	NA	NA
OLK8	0.925	0.15	1.696	0.358	2.42	0.057
OLK9	0.898	0.172	1.54	0.391	2.483	NA
OSCC1	0.979	0.096	1.959	0.278	2.41	0.289
OSCC2	0.911	0.128	1.604	0.193	3.185	0.734
OSCC3	0.88	0.14	1.564	0.327	2.393	0.127
OSCC4	0.989	0.147	1.848	0.179	2.613	0.442
OSCC5	0.991	0.091	1.596	0.123	2.772	0.396
OSCC6	0.955	0.955	1.855	0.299	2.441	0.233
OSCC7	0.983	0.099	1.579	0.148	3.2	0.265
OSCC8	0.871	0.211	1.402	0.121	2.743	0.263
OSCC9	0.884	0.104	1.694	0.225	2.605	0.238

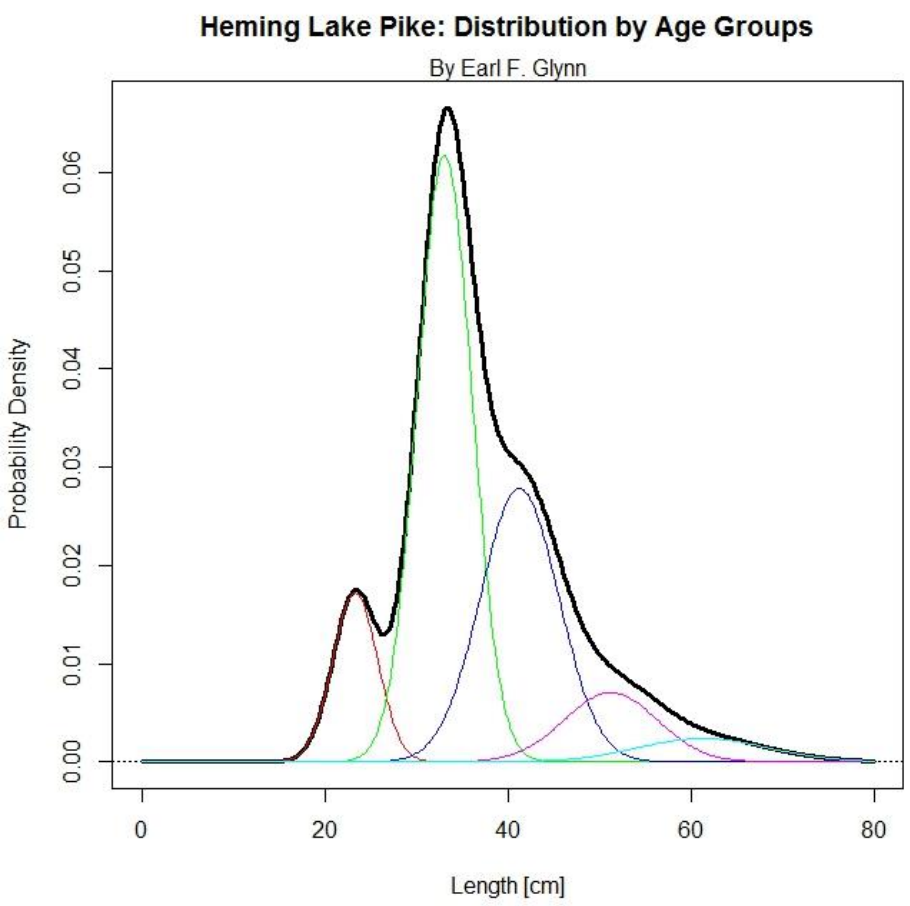

Box plot on pilot data
Missing was imputed with median or mean
0
1
2
3
Mean1
SD1
Mean2
SD2
Mean3
SD3

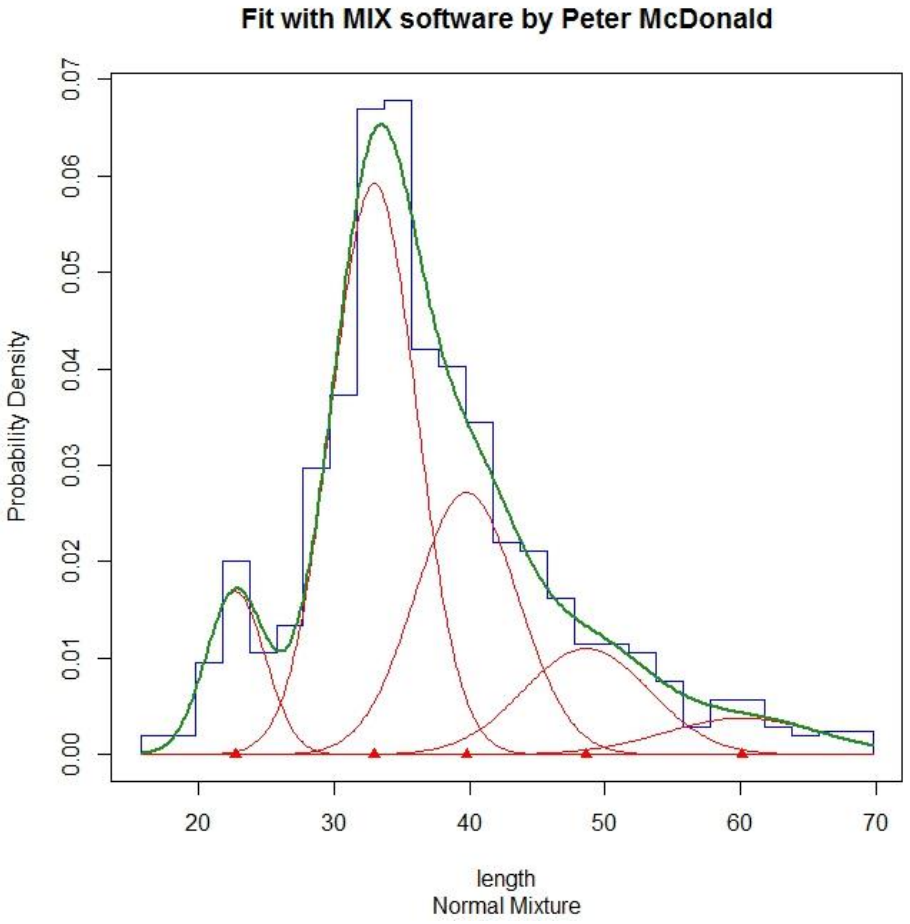
Two prediction model fitting

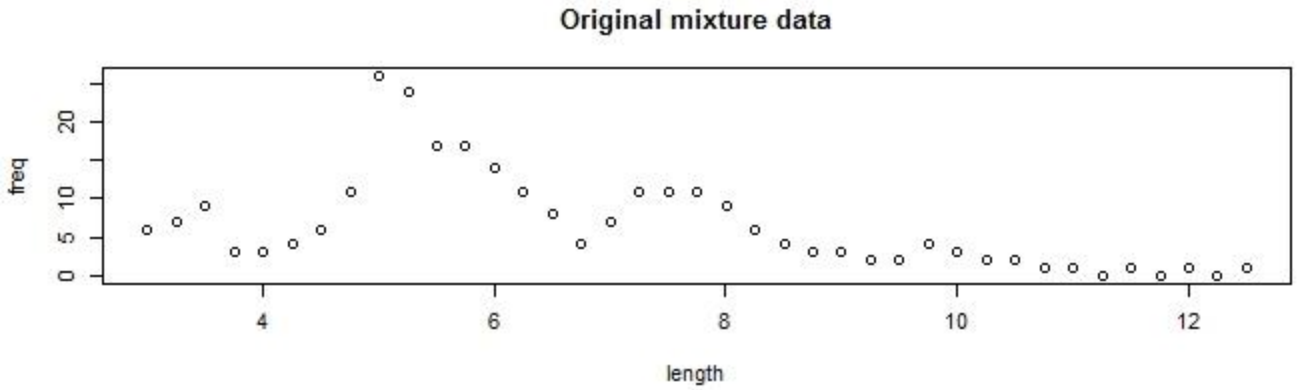
- Data was randomly separated to training and testing (7:2)
- Two models were tested
  - NaïveBayes
  - SVM
- Prediction was applied on the testing data
- And, the results

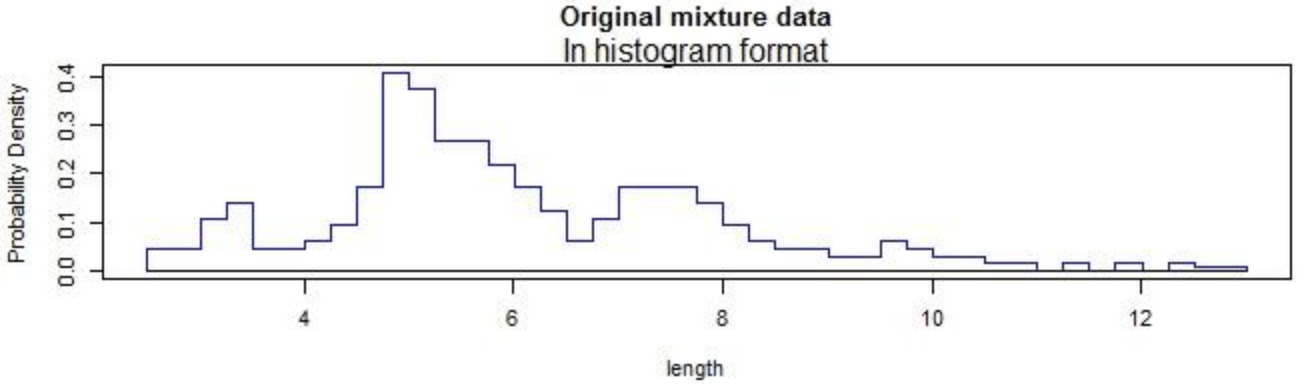
```
> table(predNB2, ans)
       ans
predNB2 norm olk oscc
   norm    1   1    1
   olk     1   1    1
   oscc    0   0    0
> table(predSVM, ans)
       ans
predSVM norm olk oscc
   norm    1   0    1
   olk     1   1    1
   oscc    0   1    0
```

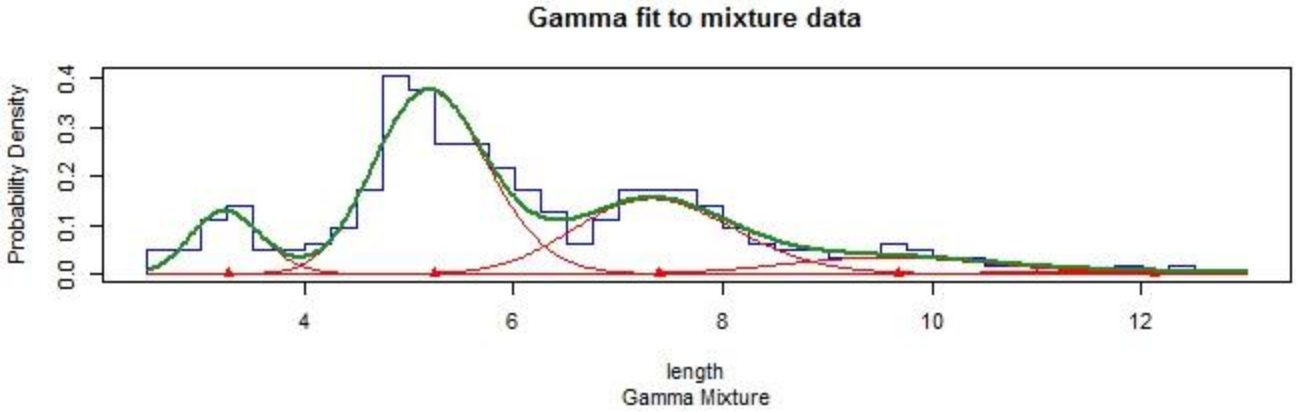
Mixture of five group data



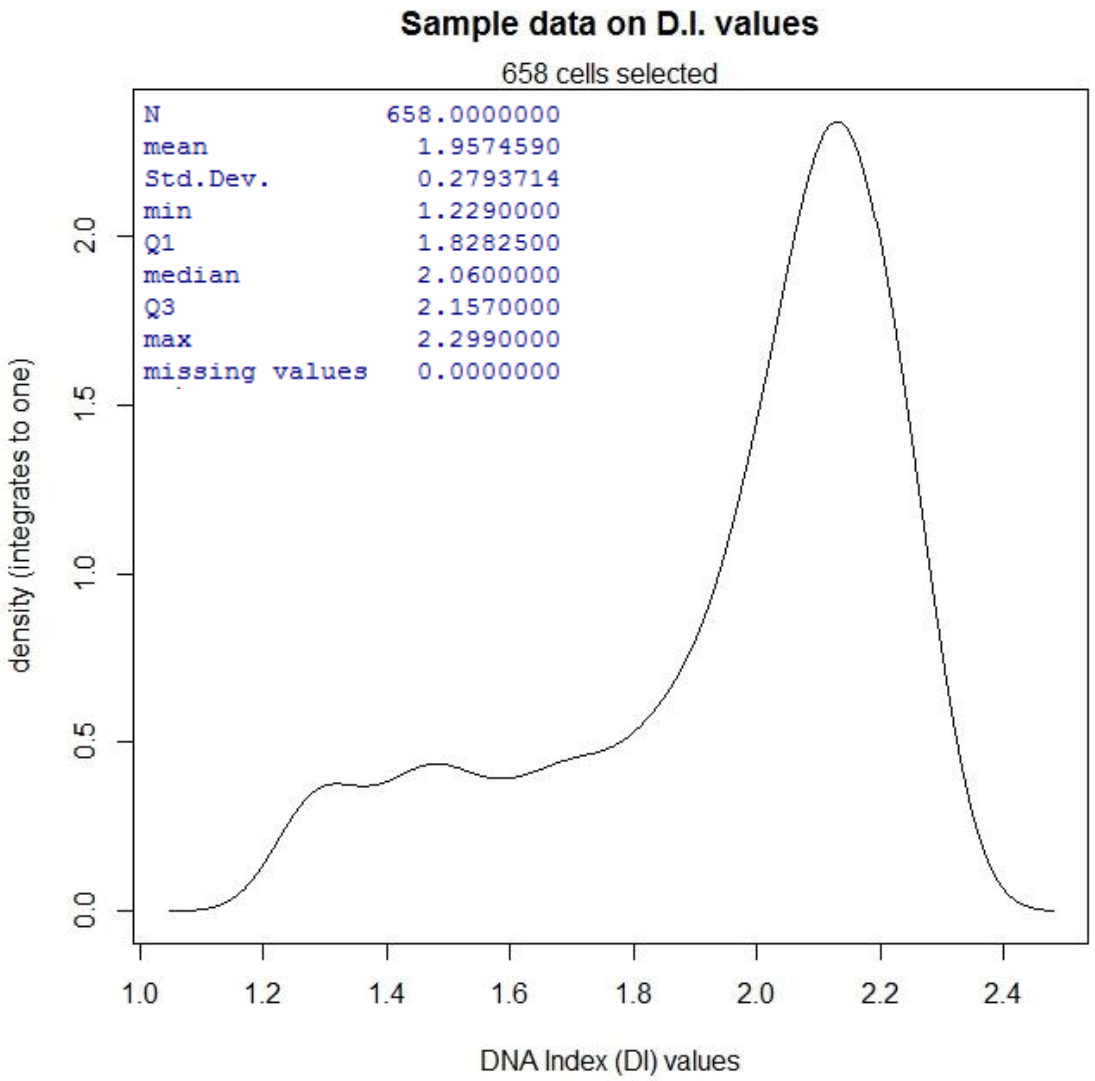



Original mixture data
freq
0
5
10
20
4
6
8
10
12
length


Original mixture data
In histogram format
Probability Density
0.0
0.1
0.2
0.3
0.4
4
6
8
10
12
length

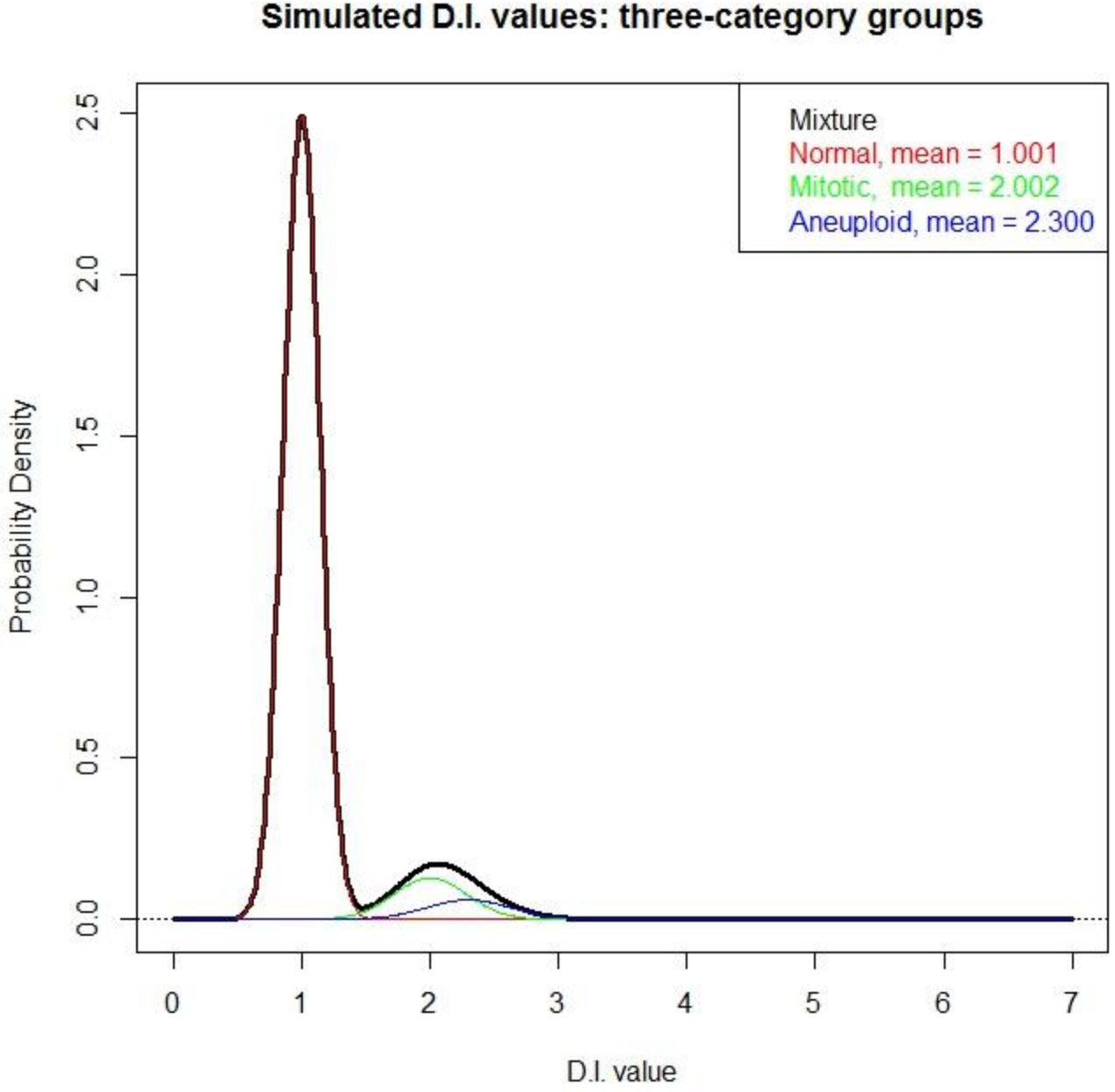

Gamma fit to mixture data
Probability Density
0.0
0.1
0.2
0.3
0.4
4
6
8
10
12
length
Gamma Mixture

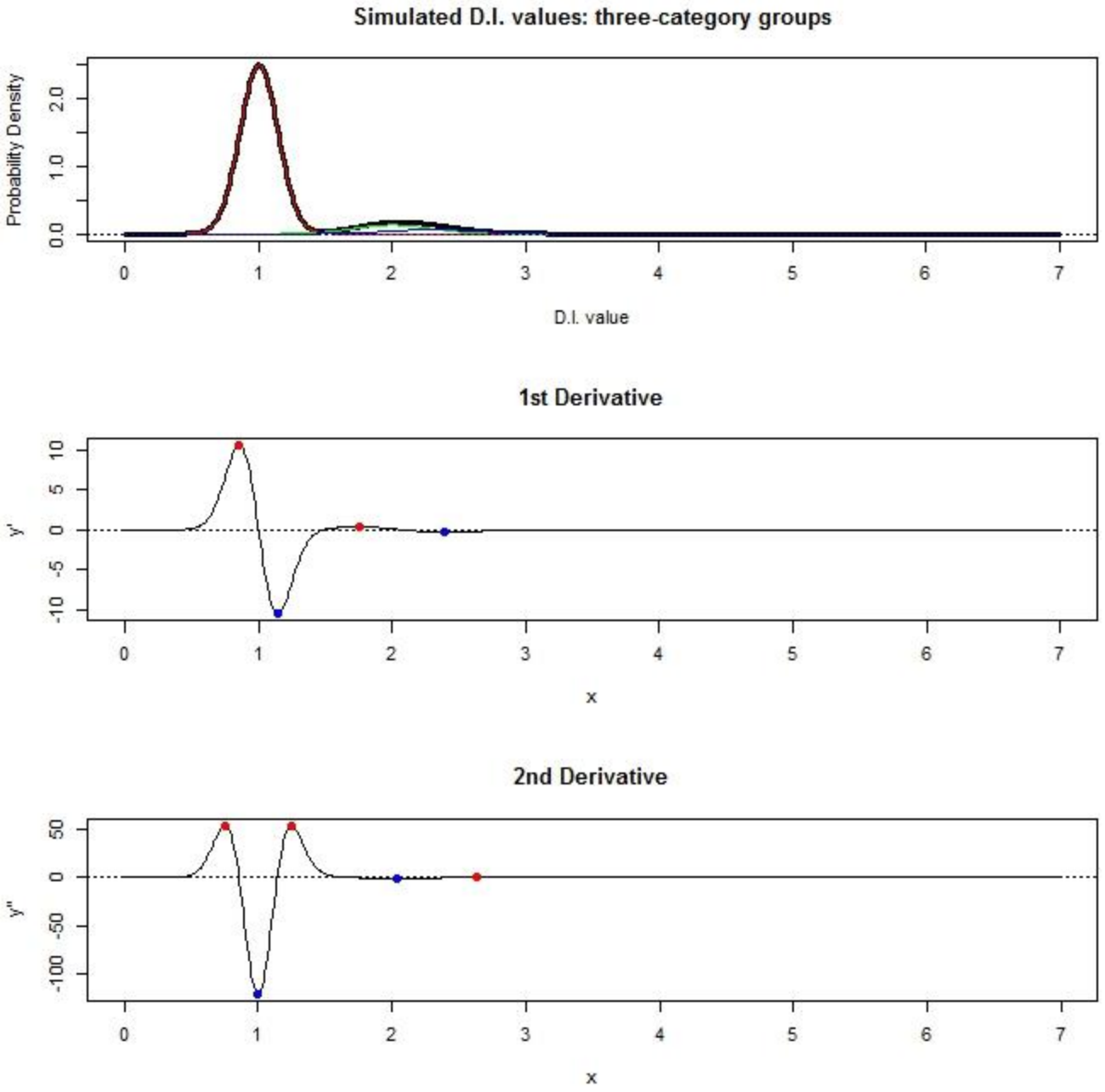
One sample data



Our first trial – step one

- Three basic clusters of cell populations
  - D.I. value: 0.9 – 1.249
  - D.I. value: 1.250 – 2.299
  - D.I. value: > 2.300
- Assuming equal C.V. across three populations, and estimated from “normal sample”:
- Three means: 1.001, 2.002, 3.003


Simulated D.I. values: three-category groups
Mixture
Normal, mean = 1.001
Mitotic, mean = 2.002
Aneuploid, mean = 2.300
Probability Density
0.0
0.5
1.0
1.5
2.0
2.5
0
1
2
3
4
5
6
7
D.I. value


Simulated D.I. values: three-category groups
Probability Density
0.0
1.0
2.0
0
1
2
3
4
5
6
7
D.I. value
1st Derivative
y'
10
5
0
-5
-10
x
2nd Derivative
y''
50
0
-50
-100
x

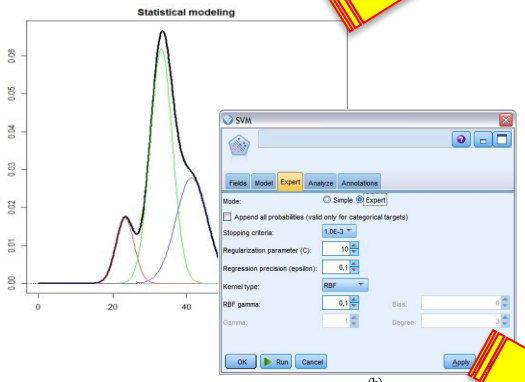
Our hurdle

- Too much information from “normal” cluster
- Possible to detect the “mitotic” stage
- Impossible to detect “aneuploidy” stage
- Need to strip out the “normal” information first, but how?
  - Don’t know
  - Don’t know
  - Don’t know

On-going effort and future goal

- Get sufficient summary statistics from ten of thousands cells (i.e. D.I. values)
- Evaluating the summary statistics
  - Reflect the mixture of distribution
  - Robust enough (enough information) to differentiation cell population
  - Capture cell population characteristics
- Building the prediction model
  - Assess model performance
  - Provide accurate diagnosis guidance
  - Constantly improve the model

Raw dataDB


Statistical modeling
SVM
Fields
Model
Expert
Analyze
Annotations
Mode:
Simple
Expert
Append all probabilities (valid only for categorical targets)
Stopping criteria:
Regularization parameter (C):
Regression precision (epsilon):
Kernel type:
RBF gamma:
Bias:
Degree:
OK
Run
Cancel
Apply

(b)


dreamstime