Robust Classification

Ruben Zamar (joint work with Mohua Podder and Will Welch)
Department of Statistics, UBC

December 12, 2011

John Tukey

- "... just which robust/resistant methods you use is not important – **what is important is that you use some...**" John. W. Tukey (1979)

PART I

BACKGROUND

AND

MOTIVATION

SNP

- SNPs are the most abundant form (90%) of genetic variability,

SNP

- SNPs are the most abundant form (90%) of genetic variability,
- SNPs are defined as DNA sequence variations that occur when a single base (A, C, G or T) in the genome is altered.

SNP

- SNPs are the most abundant form (90%) of genetic variability,
- SNPs are defined as DNA sequence variations that occur when a single base (A, C, G or T) in the genome is altered.
- Combinations of SNPs are partly responsible for
  - disease susceptibility,
  - response to illness
  - response to medical therapy
  - adverse drug reaction

SNP Genotyping

- The determination of a given person's base sequence at a specific SNP site is called SNP genotyping.

SNP Genotyping

- The determination of a given person's base sequence at a specific SNP site is called SNP genotyping.
- Many medium to high throughput genotyping techniques have been developed and tested in recent years

SNP Genotyping

- The determination of a given person's base sequence at a specific SNP site is called SNP genotyping.
- Many medium to high throughput genotyping techniques have been developed and tested in recent years
    - **Affymetrix GeneChips** (Kennedy et al., 2003)

SNP Genotyping

- The determination of a given person's base sequence at a specific SNP site is called SNP genotyping.
- Many medium to high throughput genotyping techniques have been developed and tested in recent years
  - **Affymetrix GeneChips** (Kennedy et al., 2003)
  - **Illumina's bead-array system** (Oliphant et al., 2002, Fan et al., 2006)

SNP Genotyping

- The determination of a given person's base sequence at a specific SNP site is called SNP genotyping.
- Many medium to high throughput genotyping techniques have been developed and tested in recent years
  - **Affymetrix GeneChips** (Kennedy et al., 2003)
  - **Illumina's bead-array system** (Oliphant et al., 2002, Fan et al., 2006)
- These are designed to analyze **thousands of SNPs** simultaneously

SNP Genotyping in Clinical Settings

- A challenge for the Human Genome Project is to transfer genetic knowledge to benefits society at large.

SNP Genotyping in Clinical Settings

- A challenge for the Human Genome Project is to transfer genetic knowledge to benefits society at large.
- Project: to apply SNP-related research to medical and clinical settings.

SNP Genotyping in Clinical Settings

- A challenge for the Human Genome Project is to transfer genetic knowledge to benefits society at large.
- Project: to apply SNP-related research to medical and clinical settings.
- Leading SNP genotyping technologies are ”research oriented” (expensive and relatively slow)

SNP Genotyping in Clinical Settings

- A challenge for the Human Genome Project is to transfer genetic knowledge to benefits society at large.
- Project: to apply SNP-related research to medical and clinical settings.
- Leading SNP genotyping technologies are "research oriented" (expensive and relatively slow)
- In clinical settings we need genotyping **hundreds** of SNPs simultaneously for a patient

SNP Genotyping in Clinical Settings

- A challenge for the Human Genome Project is to transfer genetic knowledge to benefits society at large.
- Project: to apply SNP-related research to medical and clinical settings.
- Leading SNP genotyping technologies are "research oriented" (expensive and relatively slow)
- In clinical settings we need genotyping **hundreds** of SNPs simultaneously for a patient
- The genotyping method should be:

SNP Genotyping in Clinical Settings

- A challenge for the Human Genome Project is to transfer genetic knowledge to benefits society at large.
- Project: to apply SNP-related research to medical and clinical settings.
- Leading SNP genotyping technologies are "research oriented" (expensive and relatively slow)
- In clinical settings we need genotyping **hundreds** of SNPs simultaneously for a patient
- The genotyping method should be:
  - rapid (e.g. couple of hours)

SNP Genotyping in Clinical Settings

- A challenge for the Human Genome Project is to transfer genetic knowledge to benefits society at large.
- Project: to apply SNP-related research to medical and clinical settings.
- Leading SNP genotyping technologies are "research oriented" (expensive and relatively slow)
- In clinical settings we need genotyping **hundreds** of SNPs simultaneously for a patient
- The genotyping method should be:
    - rapid (e.g. couple of hours)
    - accurate,

SNP Genotyping in Clinical Settings

- A challenge for the Human Genome Project is to transfer genetic knowledge to benefits society at large.
- Project: to apply SNP-related research to medical and clinical settings.
- Leading SNP genotyping technologies are "research oriented" (expensive and relatively slow)
- In clinical settings we need genotyping **hundreds** of SNPs simultaneously for a patient
- The genotyping method should be:
  - rapid (e.g. couple of hours)
  - accurate,
  - robust,

SNP Genotyping in Clinical Settings

- A challenge for the Human Genome Project is to transfer genetic knowledge to benefits society at large.
- Project: to apply SNP-related research to medical and clinical settings.
- Leading SNP genotyping technologies are ”research oriented” (expensive and relatively slow)
- In clinical settings we need genotyping **hundreds** of SNPs simultaneously for a patient
- The genotyping method should be:
  - rapid (e.g. couple of hours)
  - accurate,
  - robust,
  - cost effective

ScottTebbutt's Genotyping Approach

- Tebbut's genotyping array chip design (Tebbutt et al., 2004) is based on a redundant chemistry.

ScottTebbutt's Genotyping Approach

- Tebbut's genotyping array chip design (Tebbutt et al., 2004) is based on a redundant chemistry.
- The genotyping technology involves four probes:

ScottTebbutt's Genotyping Approach

- Tebbut's genotyping array chip design (Tebbutt et al., 2004) is based on a redundant chemistry.
- The genotyping technology involves four probes:
  - classical APEX probe, Left strand

ScottTebbutt's Genotyping Approach

- Tebbut's genotyping array chip design (Tebbutt et al., 2004) is based on a redundant chemistry.
- The genotyping technology involves four probes:
  - classical APEX probe, Left strand
  - classical APEX probe, Right strand

ScottTebbutt's Genotyping Approach

- Tebbut's genotyping array chip design (Tebbutt et al., 2004) is based on a redundant chemistry.
- The genotyping technology involves four probes:
  - classical APEX probe, Left strand
  - classical APEX probe, Right strand
  - allele-specific APEX (ASO), Left strand

ScottTebbutt's Genotyping Approach

- Tebbut's genotyping array chip design (Tebbutt et al., 2004) is based on a redundant chemistry.
- The genotyping technology involves four probes:
  - classical APEX probe, Left strand
  - classical APEX probe, Right strand
  - allele-specific APEX (ASO), Left strand
  - allele-specific APEX (ASO), Right strand

PART II

GENOTYPING MODEL

Genotyping Data

- For any given SNP we have two “expected alleles” (say alleles C and T, to fix ideas)

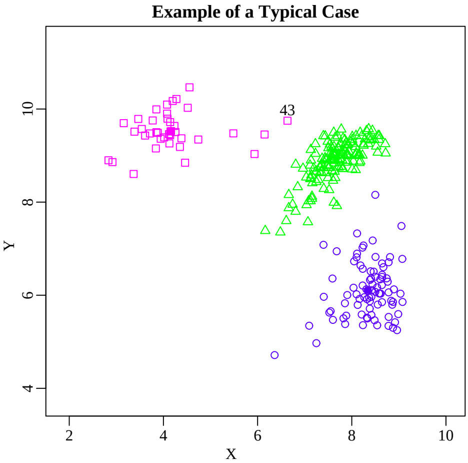
Genotyping Data

- For any given SNP we have two “expected alleles” (say alleles C and T, to fix ideas)
- From each probe, then, we get two readings:

$$X = \text{“intensity of allele C”}$$

$$Y = \text{“intensity of allele T”}$$

Genotyping Data

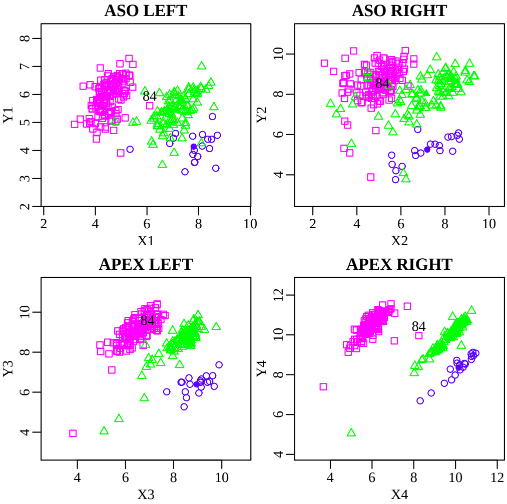


Genotyping Data (Continued)

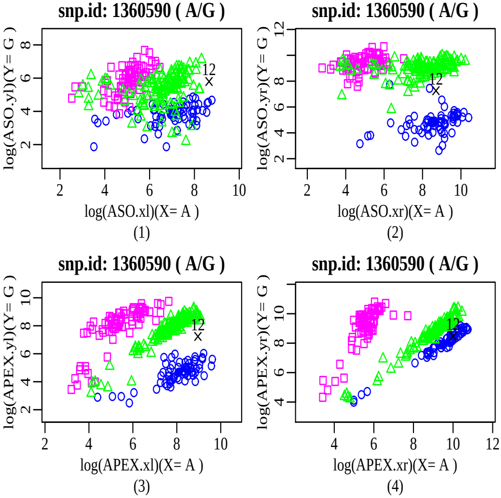
- We have a total of 4 pairs of variables (a pair from each probe)

Probe Name	Variables
ASO-Left	X_1, Y_1
ASO-Right	X_2, Y_2
APEX-Left	X_3, Y_3
APEX-Right	X_4, $Y4$

GenotypingData (Continued)



GenotypingData (Continued)



Data Sets

- To build and test the genotyping model, we have two independent data sets:

CORIEL DATA

32 Coriell DNA samples

SIRS DATA

270 DNA samples

Data Sets

- To build and test the genotyping model, we have two independent data sets:

CORIEL DATA

32 Coriell DNA samples

SIRS DATA

270 DNA samples

- CORIEL DATA: see http://coriell.umdnj.edu/; and

Data Sets

- To build and test the genotyping model, we have two independent data sets:

CORIEL DATA

32 Coriell DNA samples

SIRS DATA

270 DNA samples

- CORIEL DATA: see http://coriell.umdnj.edu/; and
- SIRS DATA: samples from systematic inflammatory response syndrome (SIRS) patients at the ICU at St. Paul's Hospital.

Data Sets

- To build and test the genotyping model, we have two independent data sets:

CORIEL DATA

32 Coriell DNA samples

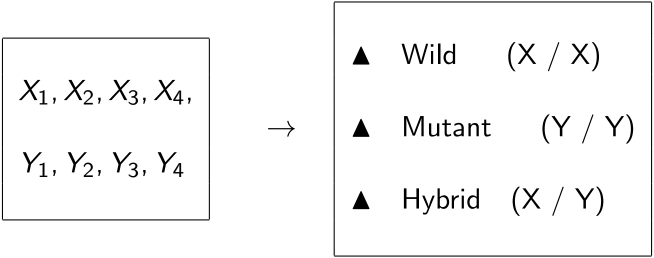
SIRS DATA

270 DNA samples

- CORIEL DATA: see http://coriell.umdnj.edu/; and
- SIRS DATA: samples from systematic inflammatory response syndrome (SIRS) patients at the ICU at St. Paul's Hospital.
- Each microarray chip has a total of 100 SNPs.

Genotyping Algorithm

Classification problem: assign each SNP/sample to one of the three possible genotypes, using the given 8 input variables



$X_1, X_2, X_3, X_4,$

Y_1, Y_2, Y_3, Y_4

$\rightarrow$

▲ Wild (X / X)

▲ Mutant (Y / Y)

▲ Hybrid (X / Y)

Building a Genotyping Model

- Conventional variables selection uses the training data to build a single (optimal) classifier.

Building a Genotyping Model

- Conventional variables selection uses the training data to build a single (optimal) classifier.
- The optimal classifier is then used to call the future test cases.

Building a Genotyping Model

- Conventional variables selection uses the training data to build a single (optimal) classifier.
- The optimal classifier is then used to call the future test cases.
- Our APEX-based genotyping platform, however, is deliberately redundant

Building a Genotyping Model

- Conventional variables selection uses the training data to build a single (optimal) classifier.
- The optimal classifier is then used to call the future test cases.
- Our APEX-based genotyping platform, however, is deliberately redundant
- The occasional failure of one or more chemistries is expected

Building a Genotyping Model

- Conventional variables selection uses the training data to build a single (optimal) classifier.
- The optimal classifier is then used to call the future test cases.
- Our APEX-based genotyping platform, however, is deliberately redundant
- The occasional failure of one or more chemistries is expected
- Therefore, occasional outliers are expected in **the training and the future data**

Our Genotyping Approach

- Our approach is to build

4 separate "base classifiers"

for each SNP.

Our Genotyping Approach

- Our approach is to build

  4 separate "base classifiers"

  for each SNP.
- Each **base classifier** uses data from a single chemistry

  ASO-LEFT
  ASO-RIGHT
  APEX-LEFT
  APEX-RIGHT

Robust Training

- Since the training data is expected to have outliers we use a robustified version of LDA, which we call RLDA

Robust Training

- Since the training data is expected to have outliers we use a robustified version of LDA, which we call RLDA

- Sample means and covariance matrices in LDA are replaced by robust S-estimates of bivariate location and scatter

Prediction

- At the **prediction stage**, the base classifiers are ensembled to call each SNP/sample

Prediction

- At the **prediction stage**, the base classifiers are ensembled to call each SNP/sample
- We use weights derived from the **"confidence"** (or lack of) associated with each base classifier

Prediction

- At the **prediction stage**, the base classifiers are ensembled to call each SNP/sample
- We use weights derived from the **"confidence"** (or lack of) associated with each base classifier
- Confidence (lack of) is assessed (dynamically) for each individual classifier and for each individual test SNP/sample.

Ensemble Using Entropy Weights

Consider the four genotype probabilities distributions and their corresponding entropies:

Chemistry	XX	YY	XY	Entropy
ASO-LEFT	p_{11}	p_{12}	p_{13}	e_1
ASO-RIGHT	p_{21}	p_{22}	p_{23}	e_2
APEX-LEFT	p_{31}	p_{32}	p_{33}	e_3
APEX-RIGHT	p_{41}	p_{42}	p_{43}	e_4
Ensembled Prob	p_1	p_2	p_3	

Ensemble Using Entropy Weights (continued)

- For $j = 1, 2, 3$ (the three differente genotypes) we set

$$p_j = \frac{p_{1j}\,(1/e_1) + p_{2j}\,(1/e_2) + p_{3j}\,(1/e_3) + p_{4j}\,(1/e_4)}{(1/e_1) + (1/e_2) + (1/e_3) + (1/e_4)}$$

Ensemble Using Entropy Weights (continued)

- For $j = 1, 2, 3$ (the three differente genotypes) we set

$$p_j = \frac{p_{1j}\,(1/e_1) + p_{2j}\,(1/e_2) + p_{3j}\,(1/e_3) + p_{4j}\,(1/e_4)}{(1/e_1) + (1/e_2) + (1/e_3) + (1/e_4)}$$

- The SNP/sample genotype is decided based on the ensembled probabilities (p_1, p_2, p_3)

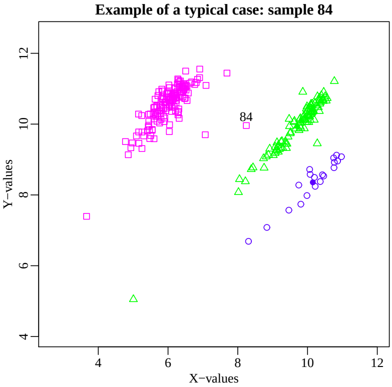
Ensemble Using Entropy Weights (continued)

- For $j = 1, 2, 3$ (the three differente genotypes) we set

$$p_j = \frac{p_{1j}\,(1/e_1) + p_{2j}\,(1/e_2) + p_{3j}\,(1/e_3) + p_{4j}\,(1/e_4)}{(1/e_1) + (1/e_2) + (1/e_3) + (1/e_4)}$$

- The SNP/sample genotype is decided based on the ensembled probabilities (p_1, p_2, p_3)
- Chemistries with less entropy are given more weight

Genotyping Case 84 - APEX-Right Base Classifier



Genotyping Case 84 - APEX-Right Base Classifier

- The genotyping results using classical LDA and RLDA are:

Method	XX	**YY**	XY
LDA	0.000	0.001	0.999
RLDA	0.000	0.0001	0.9999

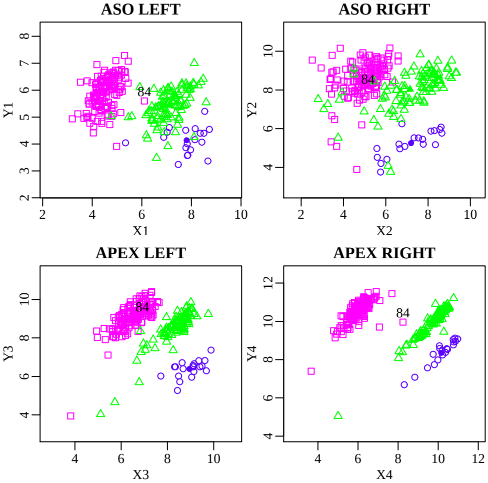
Genotyping Case 84 - APEX-Right Base Classifier

- The genotyping results using classical LDA and RLDA are:

Method	XX	YY	XY
LDA	0.000	0.001	0.999
RLDA	0.000	0.0001	0.9999

- Similar results are obtained from the ASO-Left.

Using the 4 Redundant Chemistries



Genotyping Results Using the 4 Chemistries

	Method	XX	**YY**	XY
Case 84	LAD	0.0	0.45	0.55
	RLDA	0.0	0.49	0.51

Genotyping Results Using the 4 Chemistries

-

	Method	XX	**YY**	XY
Case 84	LAD	0.0	0.45	0.55
	RLDA	0.0	0.49	0.51

- Better, but still giving the wrong genotype.

Genotyping Results Using the 4 Chemistries

-

	Method	XX	YY	XY
Case 84	LAD	0.0	0.45	0.55
	RLDA	0.0	0.49	0.51

- Better, but still giving the wrong genotype.
- **PROBLEM: ASO-Left and APEX-Right call Case 84 “YY” with high confidence!**

“Outlier-Shy” Classifier

- We need an **“outlier-shy classifier”**

“Outlier-Shy” Classifier

- We need an **“outlier-shy classifier”**
- A classifier that shows little confidence when the sample is an outlier taking as reference the training data.

“Outlier-Shy” Classifier

- We need an **“outlier-shy classifier”**
- A classifier that shows little confidence when the sample is an outlier taking as reference the training data.
- The ideal **“outlier-shy classifier”** would assign probability 1/3 to each of the three genotypes.

“Outlier-Shy” Classifier (continued)

- Instead of modelling the chemistry output (x, y) as bivariate normal we use the mixture model

$$h(x, y \mid c) = (1 - \delta) f(x, y \mid c) + \delta g(x, y)$$

“Outlier-Shy” Classifier (continued)

- Instead of modelling the chemistry output (x, y) as bivariate normal we use the mixture model

$$h(x, y \mid c) = (1 - \delta) f(x, y \mid c) + \delta g(x, y)$$

- Informative readings come from $f(x, y \mid c)$ which depends on the true genotype

$$c = XX, XY, YY$$

“Outlier-Shy” Classifier (continued)

- Instead of modelling the chemistry output (x, y) as bivariate normal we use the mixture model

$$h(x, y \mid c) = (1 - \delta) f(x, y \mid c) + \delta g(x, y)$$

- Informative readings come from $f(x, y \mid c)$ which depends on the true genotype

$$c = XX, XY, YY$$

- Non-informative readings come from $g(x, y)$

“Outlier-Shy” Classifier (continued)

- Instead of modelling the chemistry output (x, y) as bivariate normal we use the mixture model

$$h(x, y \mid c) = (1-\delta) f(x, y \mid c) + \delta g(x, y)$$

- Informative readings come from $f(x, y \mid c)$ which depends on the true genotype

$$c = XX, XY, YY$$

- Non-informative readings come from $g(x, y)$
- $0 < \delta < 0.5$ represents the probability that (x, y) is informative

Posterior Probabilities

- For each base classifier the posterior probability of $C = c$ $[c = XX, XY, YY]$ is given by

$$P\left(C = c \mid x, y\right) = \frac{p_c \; f\left(x, y \mid c\right)}{\sum_{c' \in \{XX,YY,XY\}} p_c \; f\left(x, y \mid c'\right)}$$

$$= \frac{p_c \;\; \left[\left(1 - \delta\right) f\left(x, y \mid c\right) + \delta g\left(x, y\right)\right]}{\sum_{c' \in \{XX,YY,XY\}} p_{c'} \left[\;\left(1 - \delta\right) f\left(x, y \mid c'\right) + \delta g\left(x, y\right)\right]}$$

Posterior Probabilities

- For each base classifier the posterior probability of $C = c$ $[c = XX, XY, YY]$ is given by

$$P\left(C = c \mid x, y\right) = \frac{p_c \; f\left(x, y \mid c\right)}{\sum_{c' \in \{XX,YY,XY\}} p_c \; f\left(x, y \mid c'\right)}$$

$$= \frac{p_c \;\; \left[\left(1-\delta\right) f\left(x, y \mid c\right) + \delta g\left(x, y\right)\right]}{\sum_{c' \in \{XX,YY,XY\}} p_{c'} \left[\;\left(1-\delta\right) f\left(x, y \mid c'\right) + \delta g\left(x, y\right)\right]}$$

- p_{XX}, p_{YY} and p_{XY} are the prior probabilities for the genotypes (e.g. estimated from the training data).

Some Remarks

Outlying Test Case

- Suppose that (x, y) is an outlier with respect to the training data for the three possible genotypes

Some Remarks

Outlying Test Case

- Suppose that (x, y) is an outlier with respect to the training data for the three possible genotypes
- Then $(1-\delta) f(x, y \mid c)$ is much smaller than $\delta g(x, y)$ for all $c = XX, XY, YY$

Some Remarks

Outlying Test Case

- Suppose that (x, y) is an outlier with respect to the training data for the three possible genotypes
- Then $(1-\delta) f(x, y \mid c)$ is much smaller than $\delta g(x, y)$ for all $c = XX, XY, YY$
- Therefore

$$P(C = c \mid x, y) \approx \frac{p_c}{\sum_{c' \in \{XX, YY, XY\}} p_{c'}} \approx \frac{1}{3}$$

for relatively balanced genotype probabilities.

Some Remarks (continued)

Non-Outlying Test Case

- Suppose now that (x, y) is not an outlier,

Some Remarks (continued)

Non-Outlying Test Case

- Suppose now that (x, y) is not an outlier,
- In this case δ should be small enough to not affect the posterior probability calculations.

Some Remarks (continued)

Non-Outlying Test Case

- Suppose now that (x, y) is not an outlier,
- In this case δ should be small enough to not affect the posterior probability calculations.
- On the other hand, δ should be many orders of magnitud larger than $f(x, y \mid c)$ for all c when (x, y) is an outlier.

Genotyping Case 84 - APEX-Right Base Classifier

- The genotyping results using the APEX-Right base classifier with the Gaussian and the Mixture models:

Method	XX	**YY**	XY
LDA	0.000	0.001	0.9990
RLDA	0.000	0.0001	0.9999
LDA-Mixture	0.333	0.333	0.333
RLDA-Mixture	0.333	0.333	0.333

Genotyping Case 84 - APEX-Right Base Classifier

- The genotyping results using the APEX-Right base classifier with the Gaussian and the Mixture models:

Method	XX	**YY**	XY
LDA	0.000	0.001	0.9990
RLDA	0.000	0.0001	0.9999
LDA-Mixture	0.333	0.333	0.333
RLDA-Mixture	0.333	0.333	0.333

- Similar results are obtained from the ASO-Left.

Genotyping Case 84 - Ensemble of 4 Classifiers

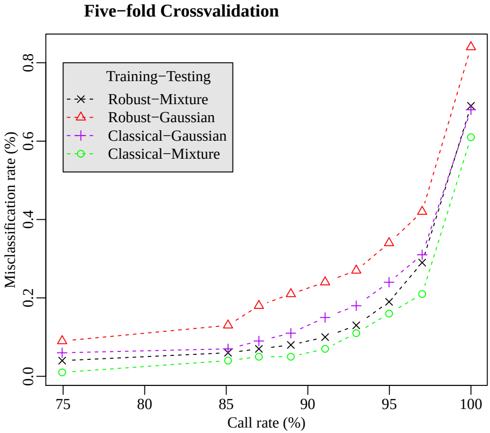
The genotyping results using the ensemble of 4 classifiers, with the Gaussian and the Mixture models:

Method	XX	YY	XY
LDA	0.000	0.45	0.55
RLDA	0.000	0.49	0.51
LDA-Mixture	0.000	0.60	0.40
RLDA-Mixture	0.000	0.66	0.34

PART III

NUMERICAL RESULTS

SIRS DATA - Five-Fold Cross Validation



Confidence Scores for APEC-Right

- Take a closer look at the behavior of each **single base classifier**

Confidence Scores for APEC-Right

- Take a closer look at the behavior of each **single base classifier**
- **Confidence Score**: posterior probabilities for the misclassified SNP/sample

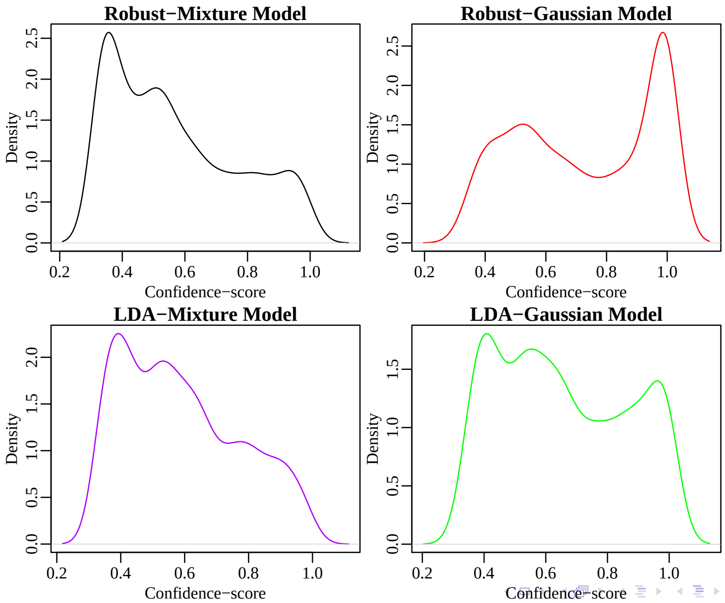
Confidence Scores for APEC-Right

- Take a closer look at the behavior of each **single base classifier**
- **Confidence Score**: posterior probabilities for the misclassified SNP/sample
- We give the results from a 5-fold-CV of SIRS data on 100 SNPs for **APEX-Right**

Confidence Scores for APEC-Right

- Take a closer look at the behavior of each **single base classifier**
- **Confidence Score**: posterior probabilities for the misclassified SNP/sample
- We give the results from a 5-fold-CV of SIRS data on 100 SNPs for **APEX-Right**
- The results for the other base classifiers are similar

Confidence Scores for APEC-Right



Simulation Results

- Generated bivariate data with approximately the same level of overlap and correlation observed in the SIRS dataset.

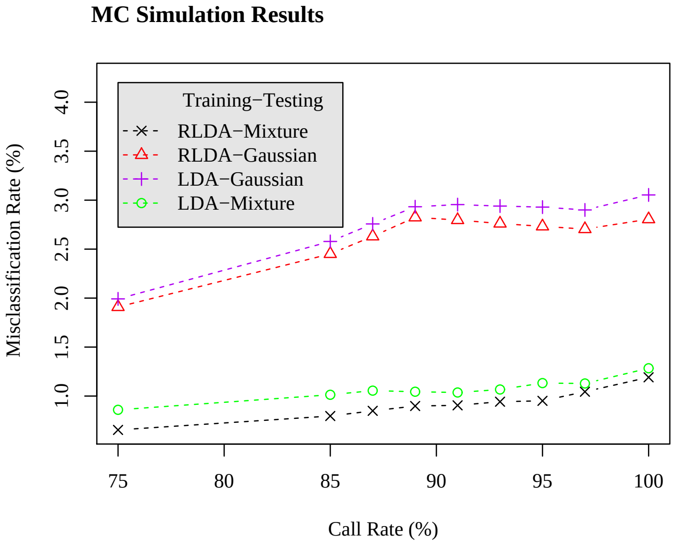
Simulation Results

- Generated bivariate data with approximately the same level of overlap and correlation observed in the SIRS dataset.
- **Training data:** added 2% of contamination (data points generated from an uniform background noise)

Simulation Results

- Generated bivariate data with approximately the same level of overlap and correlation observed in the SIRS dataset.
- **Training data:** added 2% of contamination (data points generated from an uniform background noise)
- **Testing data:** 20% probability of contamination for each test sample fed to the single base classifiers (again, data generated from an uniform background noise)

Simulation Results


MC Simulation Results
Training−Testing
RLDA−Mixture
RLDA−Gaussian
LDA−Gaussian
LDA−Mixture
Misclassification Rate (%)
Call Rate (%)