Robust Classification

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

December 12, 2011
“... 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
SNPs are the most abundant form (90%) of genetic variability,
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.
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
The determination of a given person’s base sequence at a specific SNP site is called 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.
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)
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
A challenge for the Human Genome Project is to transfer genetic knowledge to benefits society at large.
A challenge for the Human Genome Project is to transfer genetic knowledge to benefit society at large.

Project: to apply SNP-related research to medical and 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
A challenge for the Human Genome Project is to transfer genetic knowledge to benefit 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:
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)
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,
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,
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
Scott Tebbutt’s Genotyping Approach

- Tebbut’s genotyping array chip design (Tebbutt et al., 2004) is based on a redundant chemistry.
Scott Tebbutt’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:
Scott Tebbutt’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
Scott Tebbutt’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
Scott Tebbutt’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
Scott Tebbutt’s Genotyping Approach

- Tebbut’s genotyping array chip design (Tebbut 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
For any given SNP we have two “expected alleles” (say alleles C and T, to fix ideas)
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

Example of a Typical Case

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

December 12, 2011 10 / 38
We have a total of 4 pairs of variables (a pair from each probe)

<table>
<thead>
<tr>
<th>Probe Name</th>
<th>Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASO-Left</td>
<td>$X_1, Y_1$</td>
</tr>
<tr>
<td>ASO-Right</td>
<td>$X_2, Y_2$</td>
</tr>
<tr>
<td>APEX-Left</td>
<td>$X_3, Y_3$</td>
</tr>
<tr>
<td>APEX-Right</td>
<td>$X_4, Y_4$</td>
</tr>
</tbody>
</table>
GenotypingData (Continued)

snp.id: 1360590 (A/G)

log(ASO.xl)(X= A)

log(ASO.yl)(Y= G)

12

snp.id: 1360590 (A/G)

log(ASO.xr)(X= A)

log(ASO.yr)(Y= G)

12

snp.id: 1360590 (A/G)

log(APEX.xl)(X= A)

log(APEX.yl)(Y= G)

12

snp.id: 1360590 (A/G)

log(APEX.xr)(X= A)

log(APEX.yr)(Y= G)

12

Robust Classification

December 12, 2011
To build and test the genotyping model, we have two independent data sets:

- **CORIEL DATA**
  - 32 Coriell DNA samples

- **SIRS DATA**
  - 270 DNA samples
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
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.
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

\[ X_1, X_2, X_3, X_4, \]
\[ Y_1, Y_2, Y_3, Y_4 \] →

- Wild (X / X)
- Mutant (Y / Y)
- Hybrid (X / Y)
Conventional variables selection uses the training data to build a single (optimal) classifier.

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.
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.
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.
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 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
Since the training data is expected to have outliers we use a robustified version of LDA, which we call RLDA
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.
At the **prediction stage**, the base classifiers are ensembled to call each SNP/sample.
• 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
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.
Consider the four genotype probabilities distributions and their corresponding entropies:

<table>
<thead>
<tr>
<th>Chemistry</th>
<th>XX</th>
<th>YY</th>
<th>XY</th>
<th>Entropy</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASO-LEFT</td>
<td>$p_{11}$</td>
<td>$p_{12}$</td>
<td>$p_{13}$</td>
<td>$e_1$</td>
</tr>
<tr>
<td>ASO-RIGHT</td>
<td>$p_{21}$</td>
<td>$p_{22}$</td>
<td>$p_{23}$</td>
<td>$e_2$</td>
</tr>
<tr>
<td>APEX-LEFT</td>
<td>$p_{31}$</td>
<td>$p_{32}$</td>
<td>$p_{33}$</td>
<td>$e_3$</td>
</tr>
<tr>
<td>APEX-RIGHT</td>
<td>$p_{41}$</td>
<td>$p_{42}$</td>
<td>$p_{43}$</td>
<td>$e_4$</td>
</tr>
<tr>
<td>Ensembled Prob</td>
<td>$p_1$</td>
<td>$p_2$</td>
<td>$p_3$</td>
<td></td>
</tr>
</tbody>
</table>
For $j = 1, 2, 3$ (the three different genotypes) we set

$$p_j = \frac{p_{1j} \left( \frac{1}{e_1} \right) + p_{2j} \left( \frac{1}{e_2} \right) + p_{3j} \left( \frac{1}{e_3} \right) + p_{4j} \left( \frac{1}{e_4} \right)}{\left( \frac{1}{e_1} \right) + \left( \frac{1}{e_2} \right) + \left( \frac{1}{e_3} \right) + \left( \frac{1}{e_4} \right)}$$
For $j = 1, 2, 3$ (the three different genotypes) we set

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

- The SNP/sample genotype is decided based on the ensembled probabilities $(p_1, p_2, p_3)$
For $j = 1, 2, 3$ (the three different 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

Example of a typical case: sample 84

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

Robust Classification

December 12, 2011 22 / 38
The genotyping results using classical LDA and RLDA are:

<table>
<thead>
<tr>
<th>Method</th>
<th>XX</th>
<th>YY</th>
<th>XY</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDA</td>
<td>0.000</td>
<td>0.001</td>
<td>0.999</td>
</tr>
<tr>
<td>RLDA</td>
<td>0.000</td>
<td>0.0001</td>
<td>0.9999</td>
</tr>
</tbody>
</table>
The genotyping results using classical LDA and RLDA are:

<table>
<thead>
<tr>
<th>Method</th>
<th>XX</th>
<th>YY</th>
<th>XY</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDA</td>
<td>0.000</td>
<td>0.001</td>
<td>0.999</td>
</tr>
<tr>
<td>RLDA</td>
<td>0.000</td>
<td>0.0001</td>
<td>0.9999</td>
</tr>
</tbody>
</table>

Similar results are obtained from the ASO-Left.
Using the 4 Redundant Chemistries

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

Robust Classification

December 12, 2011

24 / 38
Genotyping Results Using the 4 Chemistries

<table>
<thead>
<tr>
<th>Method</th>
<th>XX</th>
<th>YY</th>
<th>XY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case 84 LAD</td>
<td>0.0</td>
<td>0.45</td>
<td>0.55</td>
</tr>
<tr>
<td>RLDA</td>
<td>0.0</td>
<td>0.49</td>
<td>0.51</td>
</tr>
</tbody>
</table>

Better, but still giving the wrong genotype.
Genotyping Results Using the 4 Chemistries

<table>
<thead>
<tr>
<th>Method</th>
<th>XX</th>
<th>YY</th>
<th>XY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case 84</td>
<td>LAD</td>
<td>0.0</td>
<td>0.45</td>
</tr>
<tr>
<td>Case 84</td>
<td>RLDA</td>
<td>0.0</td>
<td>0.49</td>
</tr>
</tbody>
</table>

- Better, but still giving the wrong genotype.
Genotyping Results Using the 4 Chemistries

Better, but still giving the wrong genotype.

**PROBLEM:** ASO-Left and APEX-Right call Case 84 “YY” with high confidence!

<table>
<thead>
<tr>
<th>Method</th>
<th>XX</th>
<th>YY</th>
<th>XY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case 84</td>
<td>0.0</td>
<td>0.45</td>
<td>0.55</td>
</tr>
<tr>
<td>RLDA</td>
<td>0.0</td>
<td>0.49</td>
<td>0.51</td>
</tr>
</tbody>
</table>
We need an "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.
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.
Instead of modelling the chemistry output \((x, y)\) as bivariate normal we use the mixture model

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

\[c = XX, XY, YY\]
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 \(g(x, y)\)
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
For each base classifier the posterior probability of $C = c$ [$c = XX, XY, YY$] is given by

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

$$= \frac{p_c [(1 - \delta) f(x, y \mid c) + \delta g(x, y)]}{\sum_{c' \in \{XX, YY, XY\}} p_{c'} [(1 - \delta) f(x, y \mid c') + \delta g(x, y)\]}$$
For each base classifier the posterior probability of \( C = c \) \([c = XX, XY, YY]\) is given by

\[
P(C = c \mid x, y) = \frac{p_c \, f(x, y \mid c)}{\sum_{c' \in \{XX, YY, XY\}} p_c \, f(x, y \mid c')}
\]

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

\( p_{XX}, p_{YY} \) 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.
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.
Suppose now that \((x, y)\) is not an outlier,
Suppose now that \((x, y)\) is not an outlier,

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

In this case \(\delta\) should be small enough to not affect the posterior probability calculations.

On the other hand, \(\delta\) should be many orders of magnitude 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:

<table>
<thead>
<tr>
<th>Method</th>
<th>XX</th>
<th>YY</th>
<th>XY</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDA</td>
<td>0.000</td>
<td>0.001</td>
<td>0.9990</td>
</tr>
<tr>
<td>RLDA</td>
<td>0.000</td>
<td>0.0001</td>
<td>0.9999</td>
</tr>
<tr>
<td>LDA-Mixture</td>
<td>0.333</td>
<td>0.333</td>
<td>0.333</td>
</tr>
<tr>
<td>RLDA-Mixture</td>
<td>0.333</td>
<td>0.333</td>
<td>0.333</td>
</tr>
</tbody>
</table>
The genotyping results using the APEX-Right base classifier with the Gaussian and the Mixture models:

<table>
<thead>
<tr>
<th>Method</th>
<th>XX</th>
<th>YY</th>
<th>XY</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDA</td>
<td>0.000</td>
<td>0.001</td>
<td>0.9990</td>
</tr>
<tr>
<td>RLDA</td>
<td>0.000</td>
<td>0.0001</td>
<td>0.9999</td>
</tr>
<tr>
<td>LDA-Mixture</td>
<td>0.333</td>
<td>0.333</td>
<td>0.333</td>
</tr>
<tr>
<td>RLDA-Mixture</td>
<td>0.333</td>
<td>0.333</td>
<td>0.333</td>
</tr>
</tbody>
</table>

Similar results are obtained from the ASO-Left.
The genotyping results using the ensemble of 4 classifiers, with the Gaussian and the Mixture models:

<table>
<thead>
<tr>
<th>Method</th>
<th>XX</th>
<th>YY</th>
<th>XY</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDA</td>
<td>0.000</td>
<td>0.45</td>
<td>0.55</td>
</tr>
<tr>
<td>RLDA</td>
<td>0.000</td>
<td>0.49</td>
<td>0.51</td>
</tr>
<tr>
<td>LDA-Mixture</td>
<td>0.000</td>
<td>0.60</td>
<td>0.40</td>
</tr>
<tr>
<td>RLDA-Mixture</td>
<td>0.000</td>
<td>0.66</td>
<td>0.34</td>
</tr>
</tbody>
</table>
PART III

NUMERICAL RESULTS
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
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

Robust–Mixture Model

Robust–Gaussian Model

LDA–Mixture Model

LDA–Gaussian Model

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

Robust Classification

December 12, 2011
• 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

Graph showing the relationship between Call Rate (%) and Misclassification Rate (%). The graph includes lines for different models:
- RLDA–Mixture
- RLDA–Gaussian
- LDA–Gaussian
- LDA–Mixture

The x-axis represents the Call Rate (%) ranging from 75 to 100, and the y-axis represents the Misclassification Rate (%) ranging from 1.0 to 4.0.