Genotype imputation based on discriminant and cluster analysis

Medhat Mahmoud
Theo Meuwissen
Thore Egeland
Department of Animal and Aquaculture Sciences
Norwegian University of Life Sciences
Ås, Norway

Overview

What Is Imputation?

- Is the replacement of a missing or incorrectly reported item using logical edits or statistical procedures

- In other words, Imputation replaces a missing or incorrect data item with an "educated guess."

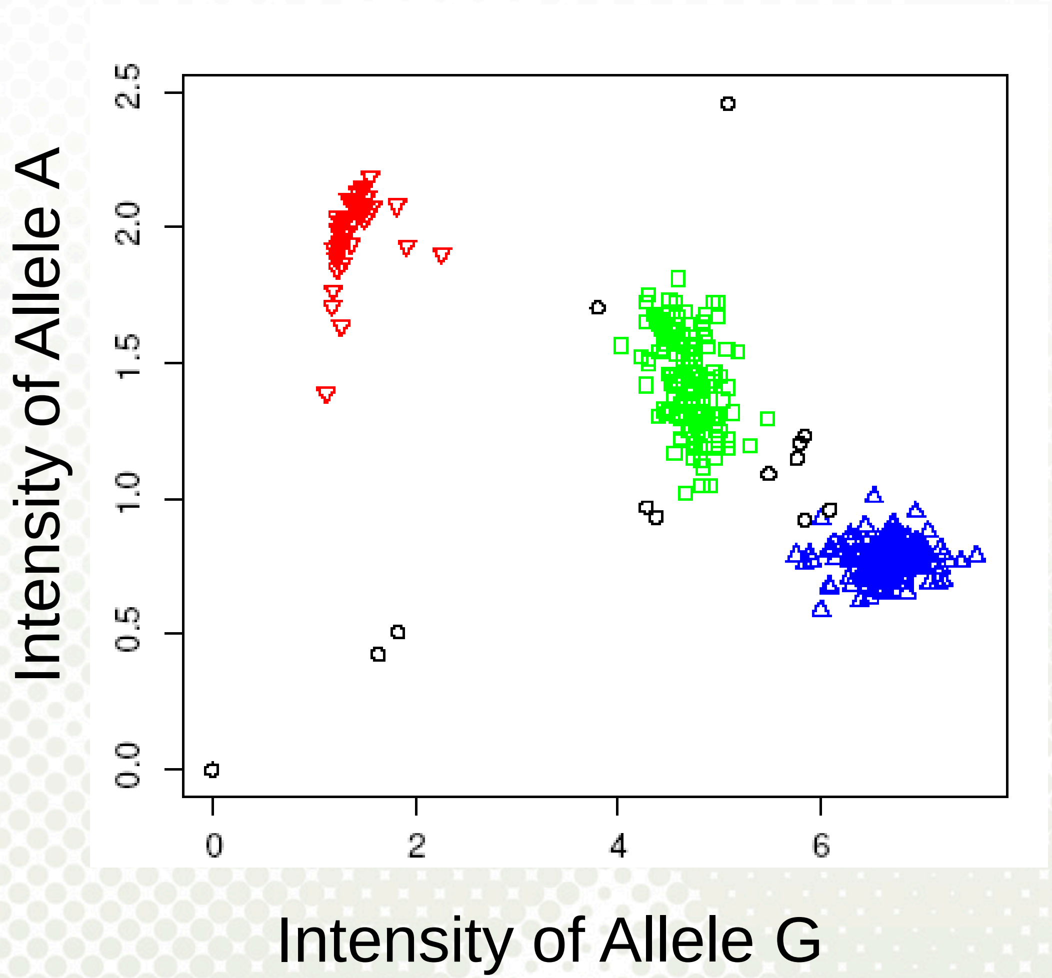
Department of Animal and Aquaculture Sciences
UNIVERSITETET FOR MILJØ- OG BIOVITENSKAP

Genotype imputation

- **Imputation** of genotypes at **un-typed** SNP loci
  - Powerful technique for increasing the power of association studies
  - Typed markers in conjunction with **catalogs of SNP variation** (e.g. HapMap) ➔ **predictors** for SNP not present on the array

- **Challenge:** Optimally combining the multi-locus information from current + multi-locus variation from HapMap

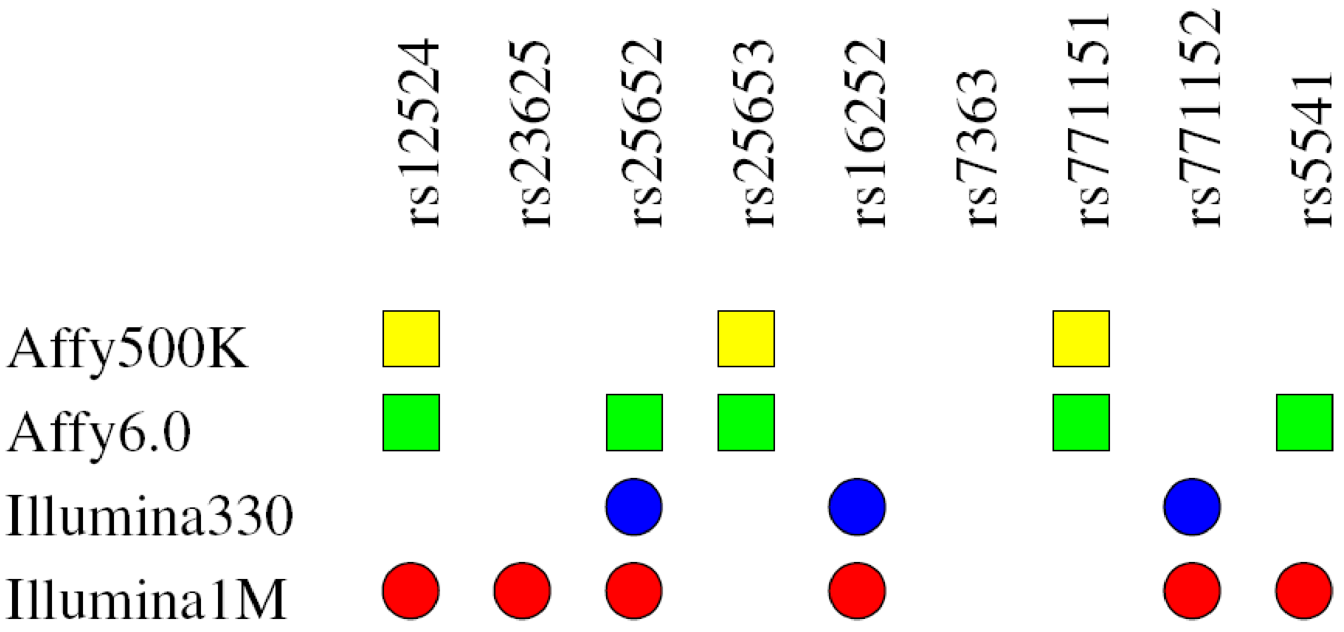
Genotypes are called with varying uncertainty



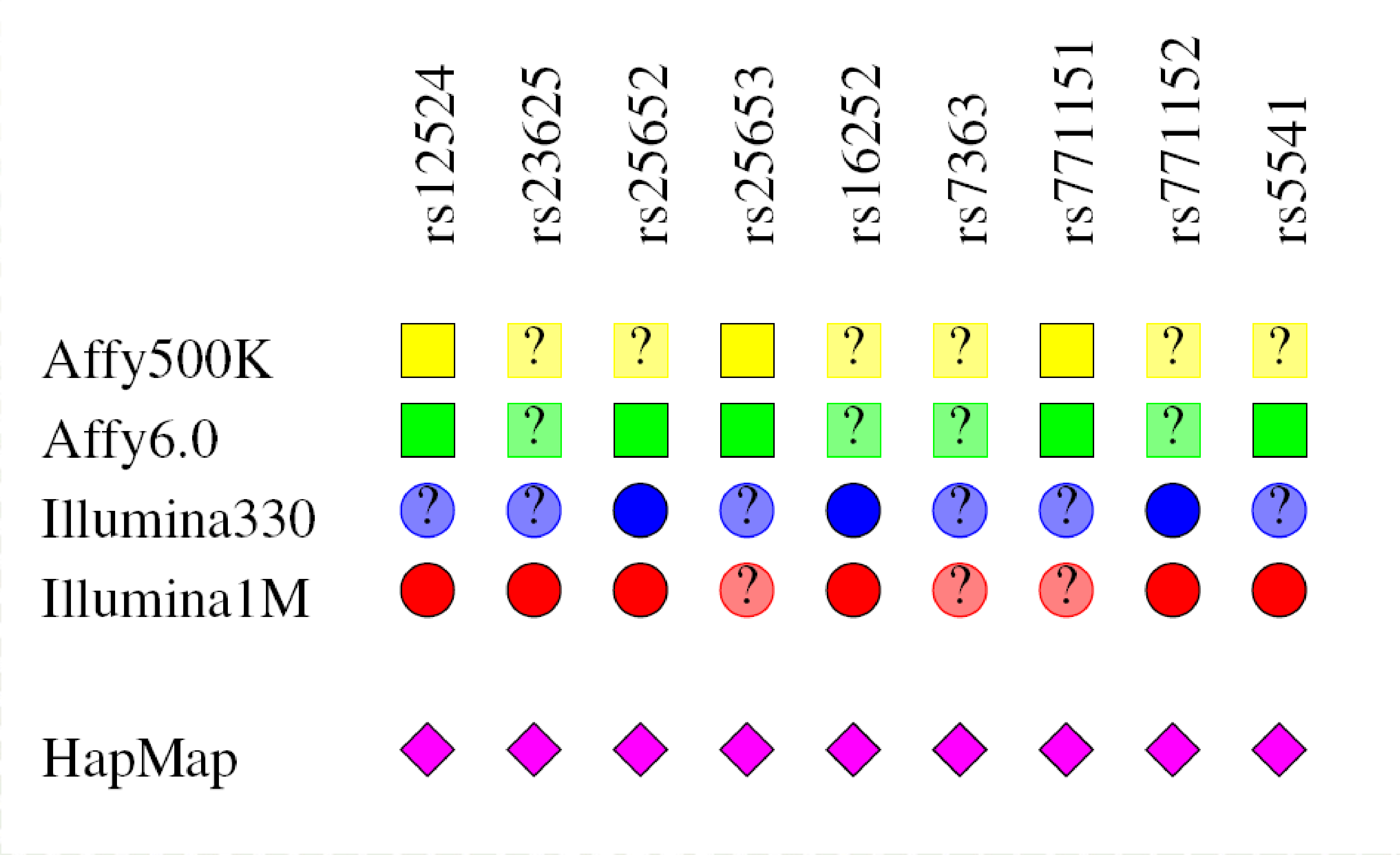
▽ = AA □ = AG △ = GG ○ = not called

Department of Animal and Aquaculture Sciences
UNIVERSITETET FOR MILJØ- OG BIOVITENSKAP

Some Genotypes are missing at all …

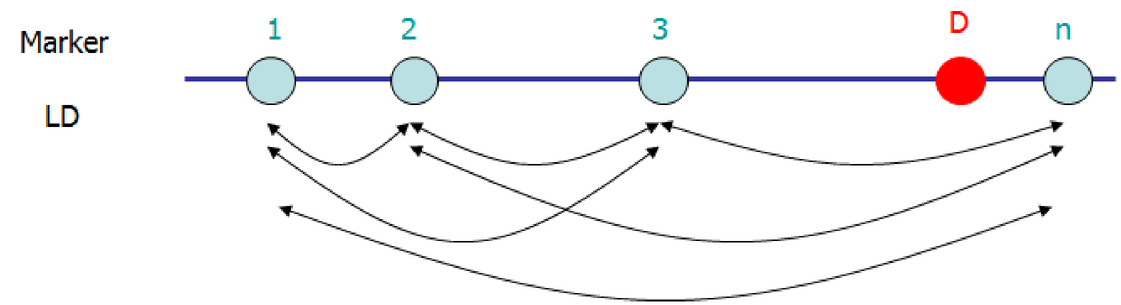


... but are imputed with different uncertainties



Department of Animal and Aquaculture Sciences
UNIVERSITETET FOR MILJØ- OG BIOVITENSKAP

Or using Linkage Disequilibrium between markers !



Markers close together on chromosomes are often transmitted together, yielding a non-zero correlation between the alleles.

Department of Animal and Aquaculture Sciences
UNIVERSITETET FOR MILJØ- OG BIOVITENSKAP

Imputation programs are available ...

- IMPUTE
  - Developed by Jonathan Marchini
  - Nature Genetics, Advance online publication
  - http://www.stats.ox.ac.uk/~marchini/#software

- Mach 1.0, Markov Chain Haplotyping
  - Developed by Goncalo Abecasis
  - http://www.sph.umich.edu/csg/abecasis/MACH/

- BEAGLE 3.3.2
  - Developed by Brian L. Browning
  - http://faculty.washington.edu/browning/beagle/beagle.html

Aim of the study

- Testing the performance of linear discriminant and clustering analysis in SNP imputation, in 5 different situations.
  1. with different Haplotype block sizes in low Linkage disequilibrium genome region.
  2. with different Haplotype block sizes in High Linkage disequilibrium genome region.
  3. In different levels of Minor allele frequency genome regions (MAF).
  4. In different levels Marker density regions (HD, LD).
  5. with different Reference sample sizes (n).

Department of Animal and Aquaculture Sciences
UNIVERSITETET FOR MILJØ- OG BIOVITENSKAP

- Introduction
  - Imputation and Multiple imputation
  - Genotype imputation
  - Aim of the study
- Materials and Simulations
- Methods
  - Linear discriminant analysis
  - Clustering analysis
- Validation
- Results
- Discussion and Conclusion
- Questions

Department of Animal and Aquaculture Sciences
UNIVERSITETET FOR MILJØ- OG BIOVITENSKAP

Materials and Simulations

- Many datasets have been simulated for this study (See Table1)

Dataset	Test	Correlation	MAF %	No. haplotypes	No. SNP
1	No. of SNPs in **(LLD)** region	0.2	49	1000	Vary
2	No. of SNPs in **(HLD)** region	0.8	49	1000	Vary
3	Minor allele frequency **(MAF)**	0.2	Vary	1000	10
4	Marker density **(MD)**	Vary	49	1000	10
5	Reference sample size **(n)**	0.2	10	1000	10

Dataset 1 and 2:

- Simulated to investigate the effect of the different numbers of SNPs (markers) in each haplotype block in imputation Accuracy rate, in a regions of **low and High linkage disequilibrium**.

Test 2: using 10 SNPs

Test 1: using 5 SNPs

1000 Haplotype blocks, 500 as a training dataset and 500 blocks as a tested dataset

	snp1	...	snp45	snp46	snp47	snp48	snp49	snp50	snp51	snp52	snp53	snp54	snp55	...	snp K
Haplotype 1	2		2	1	1	1	2	?	1	2	1	1	1		2
Haplotype 2	1		2	1	1	2	1	?	1	2	2	2	2		1
								Un-typed SNP							
Haplotype n	2	2	1	1	2	2	2	?	1	2	1	1	2		1

Figure 2.1: small example to illustrating Dataset 1

Dataset 3

- Simulated to investigate the effects of Minor allele frequency (MAF) of un-typed SNPs in imputation accuracy rate.

part 1: 100 haplotypes, MAF=0.01

part 2: 100 haplotypes, MAF=0.02

parts 3, 4, 5, 6, 7, 8 and part 9

part 10: 100 haplotypes, MAF=0.49

	snp1	snp2	snp3	snp4	snp5	snp6	snp7	snp8	snp9	snp10
Haplotype 1	2	1	1	1	?	1	2	1	1	1
Haplotype 2	2	1	1	2	?	1	2	2	2	2
Haplotype n	1	1	2	2	?	1	2	1	1	2

Figure 2.2: small example to illustrating Dataset 2.

Dataset 4

- - Simulated to investigate the effects of marker density (MD) in imputation accuracy rate.
- - In this case we duplicate the Dataset to 10 different datasets, each one varied from the others in their correlation between SNPs, but constants with other parameters.
- - Each dataset Consisted of 1000 haplotype blocks (1000 rows, 500 haplotypes as a training dataset and 500 haplotypes as a test dataset), with 10 SNPs in each haplotype block and MAF = 0.50.

Dataset 5

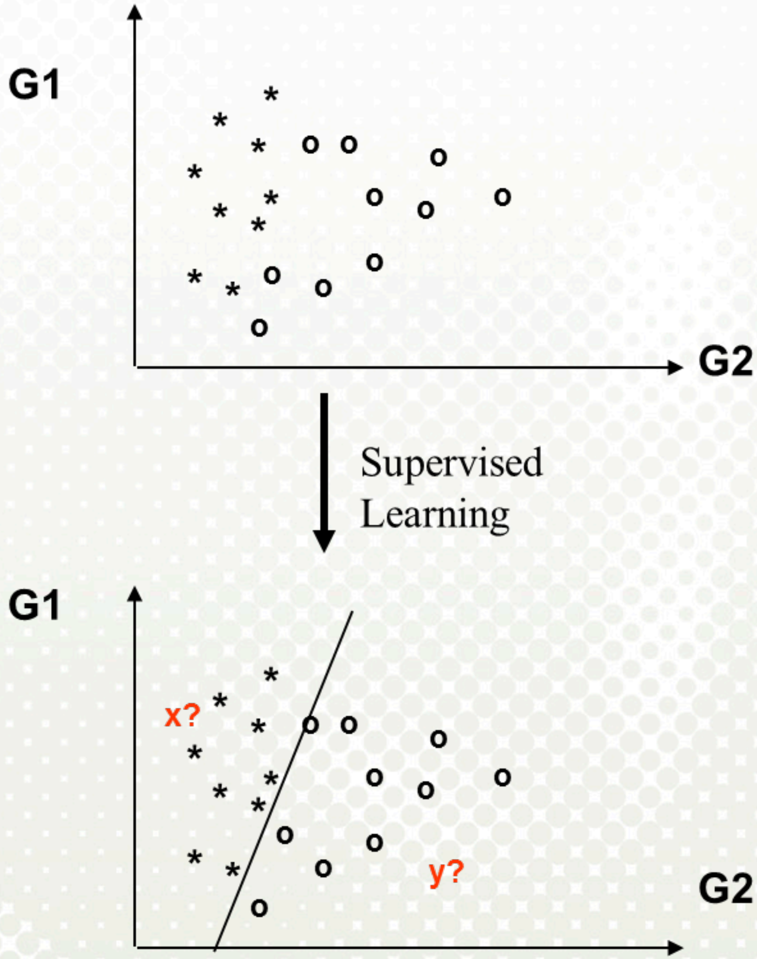
- **-** Simulated to investigate the effects of reference sample size (n) in imputation accuracy rate.
- - In this case to investigate the effects of reference sample size (n), we divided our dataset into 9 sup-datasets: 9 training datasets and 9 test datasets.
- **Sup-datasets 1:** consisted of 100 haplotypes as training-dataset and the rest 900 haplotypes as test-dataset.
- **Sup-datasets 2:** consisted of 200 haplotypes as training-dataset and the rest 800 haplotypes as test-dataset.
  and so on until the sup-datasets 9
- **Sup-datasets 9:** consisted of 900 haplotypes as training-dataset and the rest 100 haplotypes as test-dataset.
- - Each Test consisted of 1000 haplotypes and 10 SNPs, with correlation between SNPs = 0.20 and MAF =0.10.

Overview

- Introduction
  - Imputation and Multiple imputation
  - Genotype imputation
  - Aim of the study
- Materials and Simulations
- Methods
  - Linear discriminant analysis
  - Clustering analysis
- Validation
- Results
- Discussion and Conclusion
- Questions

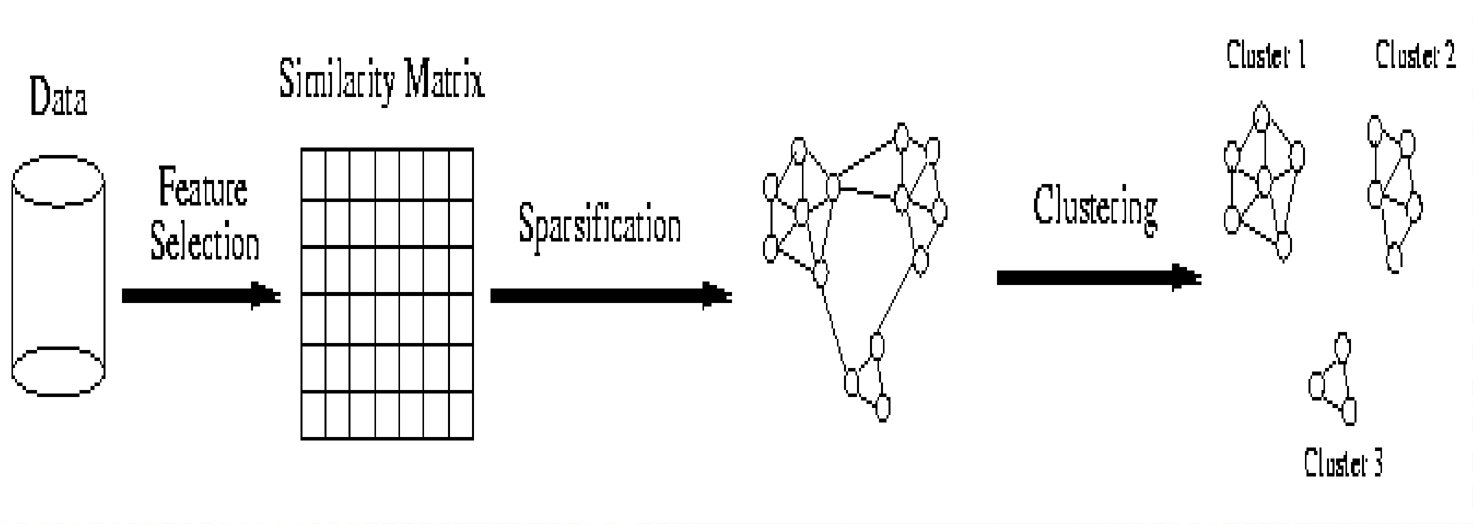
Classification

- In classification you do have a class label (o and x), each defined in terms of G1 and G2 values.
- You are trying to find a model that splits the data elements into their existing classes
- You then assume that this model can be used to assign new data points x and y to the right class



Department of Animal and Aquaculture Sciences
UNIVERSITETET FOR MILJØ- OG BIOVITENSKAP

Sparsification in the Clustering Process

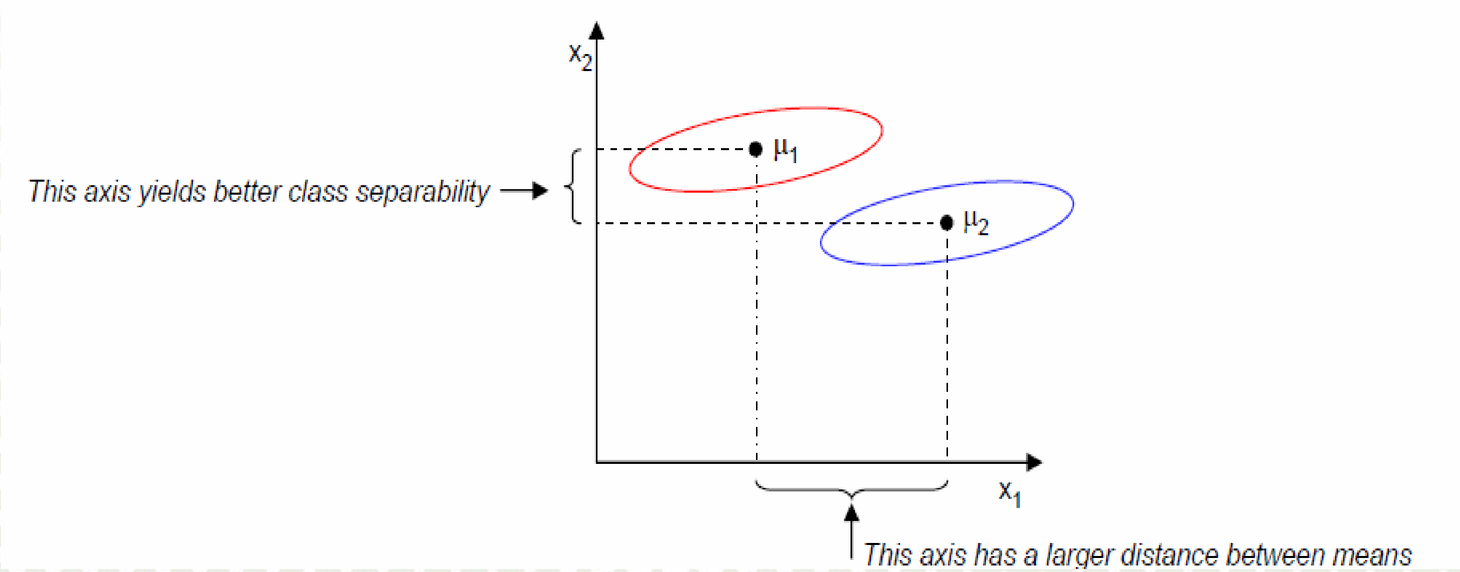


Methods

- Linear Discriminant Analysis
  - Maximum Likelihood Discriminant Rule.
    - Quadratic discriminant analysis (QDA).
    - Linear discriminant analysis (LDA, equivalent to FLDA for K=2).
    - Diagnal quadratic discriminant analysis (DQDA).
    - Diagnal linear discriminant analysis (DLDA).
  - Fisher Linear Discriminant Analysis.
- Clustering Analysis
  - Classification and Regression Tree (CART).
  - Aggregating & Bagging.
  - Nearest Neighbor Classification.

Fisher Linear Discriminant Analysis

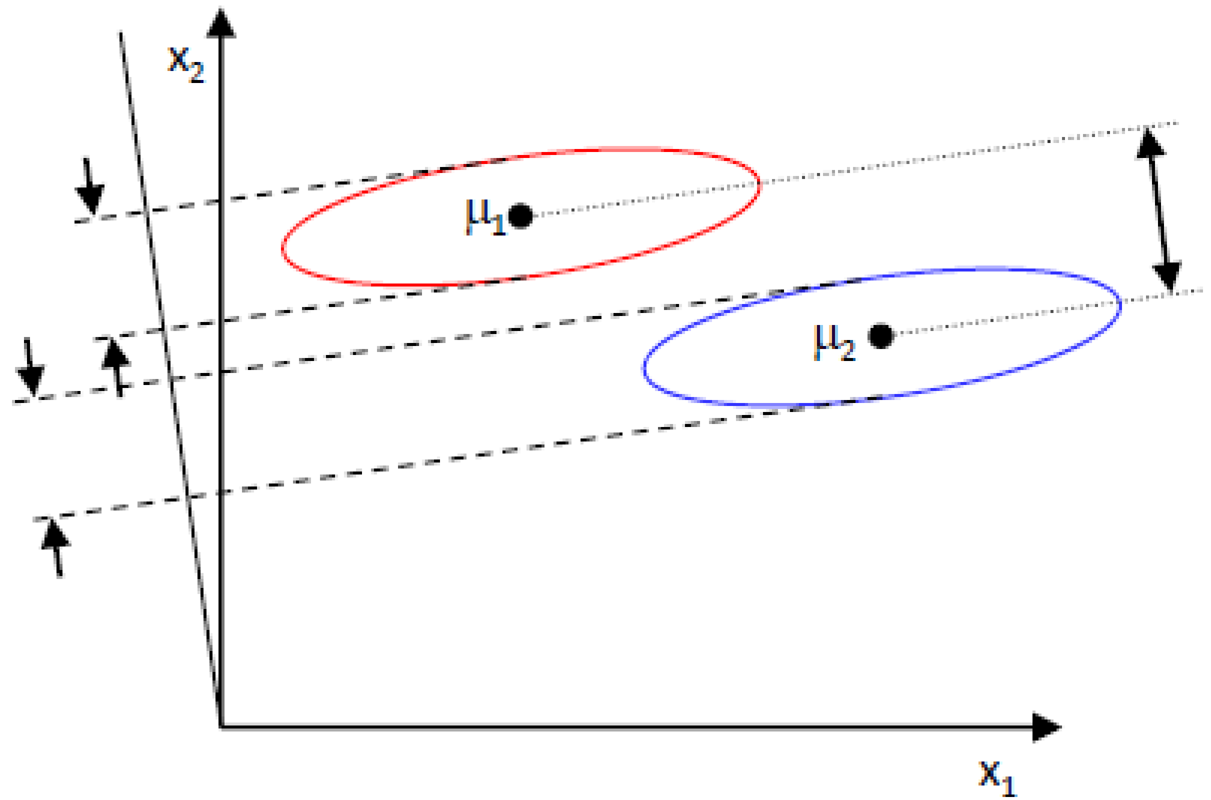
- In a two-class classification problem, given n samples in a d-dimensional feature space. n1 in class 1 and n2 in class 2.
- Goal: to find a vector w, and project the n samples on the axis y=w'x, so that the projected samples are well separated.



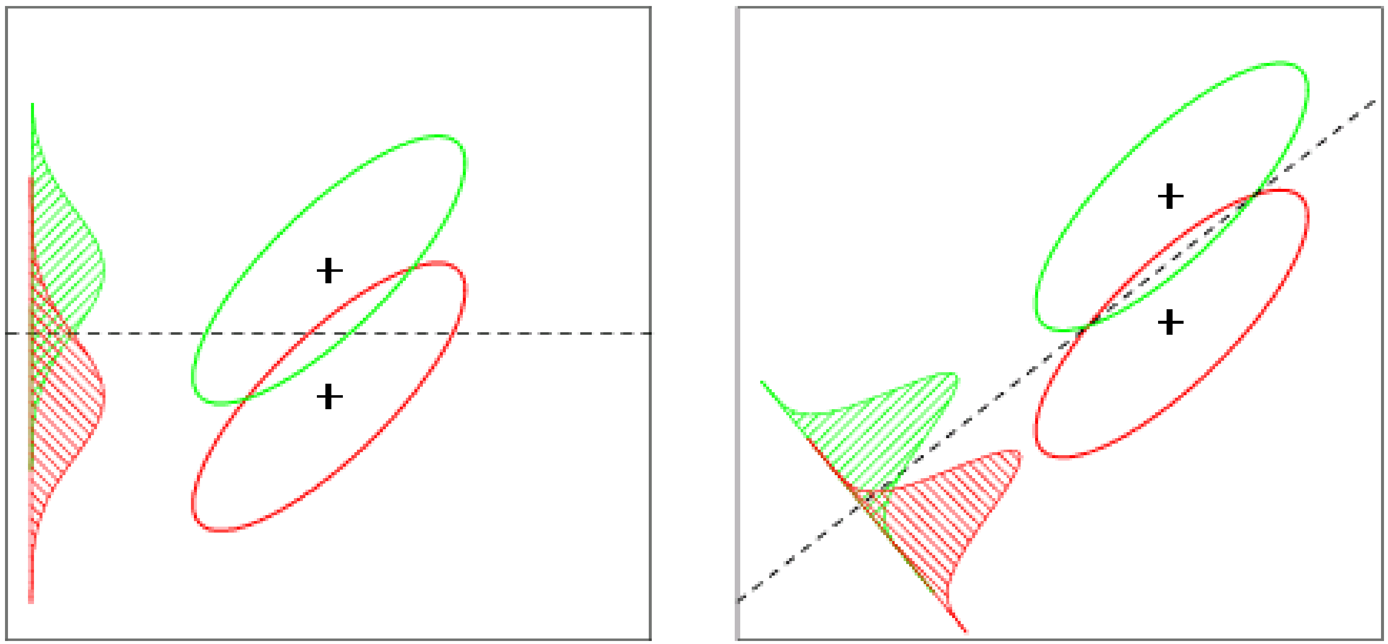
Department of Animal and Aquaculture Sciences

UNIVERSITETET FOR MILJØ- OG BIOVITENSKAP

Fisher Linear Discriminant Analysis



Department of Animal and Aquaculture Sciences
UNIVERSITETET FOR MILJØ- OG BIOVITENSKAP



Fisher Linear Discriminant Analysis

Let us assume that any given variables of SNPs (in a given haplotype block) can be described by vector X of p characteristics $(x_1, x_2, \ldots, xp)$, that can be measured ($x_{1=1}$ for major allele and $x_{1=2}$ for minor allele). The linear discriminant analysis procedure finds a linear combination of the measures (called the linear discriminant function or LDF), that provides maximum discrimination between major alleles (class 1 or ' π_1') and minor alleles (class 2 or 'π_2').

Fisher Linear Discriminant Analysis

- $Z = \sum_{i=1}^{p} \alpha_i x_i$ …….. (LDF)

- The LDF classifies *X* into class 1 if *Z* > *c* and into class 2 if *Z* < *c*. The vector of coefficients $(\alpha_1, \alpha_2, \ldots, \alpha_p)$ and threshold constant *c* were derived from the training set by maximizing the ratio of between-class variation of z to the within-class variation (Afifi and Azen, 1979)

Department of Animal and Aquaculture Sciences
UNIVERSITETET FOR MILJØ- OG BIOVITENSKAP

Fisher Linear Discriminant Analysis

- $\vec{a} = s^{-1}(\vec{m}_1 - \vec{m}_2)$

- And

- $\vec{c} = \vec{a}(\vec{m}_1 + \vec{m}_2)/2,$

- Where $\vec{m}_i$ are the sample mean vectors of characteristics for class 1 and class 2, respectively; s is pooled covariance matrix of characteristics

- $S = \frac{1}{n_1+n_2-2}(s_1 + s_2),$

Department of Animal and Aquaculture Sciences
UNIVERSITETET FOR MILJØ- OG BIOVITENSKAP

Implementation of LDA in SNP imputation

Table 3.2.2: Training dataset

Hap.	SNP1	SNP2	SNP3	SNP4	SNP5
1	2	1	2	1	1
2	1	1	1	2	2
3	2	2	1	1	2
4	1	1	2	1	1
5	1	1	2	2	2

Table 3.2.3: Test dataset

Hap.	SNP1	SNP2	SNP3	SNP4	SNP5
6	1	1	?	1	2
7	2	2	?	1	2

R commands

```
lda(SNP3 ~ SNP1 + SNP2 + SNP4 + SNP5, data = Training)
Coefficients of linear discriminants: LD1
SNP1  1.939638e-16
SNP2 -1.718108e+00
SNP4 -1.145405e-01
SNP5 -1.489027e+00
```

So the LDA model should be

SNP3 ≈μ+SNP1 (1.939638e-16) +SNP2 (-1.718108e+00) +SNP4 (-1.145405e-01) + SNP5 (-1.489027e+00) + e

e: error

Now, in order to identify the missing SNP number 3 in the Test dataset, e.g. haplotype number 6

```
predict(DAModel.5, data.frame('SNP1'=1, 'SNP2'=1, 'SNP4'=1,
'SNP5'=2))
$class
[1] 2
```

So the SNP3 in haplotype 6 (record no. 6) expected to = 2 (major allele class).

LDA vs. Logistic Regression

- LDA (Generative model)
  - Easier to train, low variance, more efficient if model is correct
  - Higher asymptotic error, but converges faster
- Logistic Regression (Discriminative model)
  - Ignores marginal density information $\Pr(X)$
  - Harder to train, robust to uncertainty about the data generation process
  - Lower asymptotic error, but converges more slowly

LDA vs. Principal component analysis.

- A tendency in the computer vision community to prefer LDA over PCA

Because LDA deals directly with discrimination between classes while PCA does not pay attention to the underlying class structure.

Fisher Linear Discriminant Analysis

M.Barnard. The secular variations of skull characters in four series of egyptian skulls. Annals of Eugenics, 6:352-371, 1935.

R.A.Fisher. The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7:179-188, 1936.

A. Martinez, A. Kak, "PCA versus LDA", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 2, pp. 228-233, 2001.

Department of Animal and Aquaculture Sciences
UNIVERSITETET FOR MILJØ- OG BIOVITENSKAP

Overview

Department of Animal and Aquaculture Sciences
UNIVERSITETET FOR MILJØ- OG BIOVITENSKAP

Nearest Neighbor

- Based on a measure of distance between observations (e.g. Euclidean distance or one minus correlation *10).

$$d(x,y) = ||x-y|| = \sqrt{(x-y)\cdot(x-y)} = \left(\sum_{i=1}^{m}(x_i-y_i)^2\right)^{1/2}$$

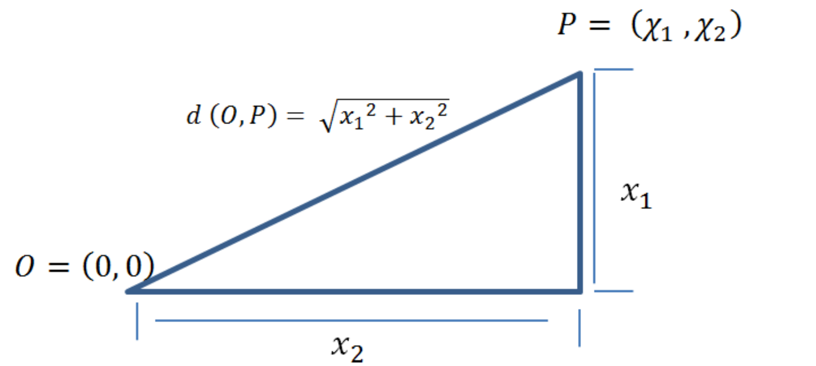


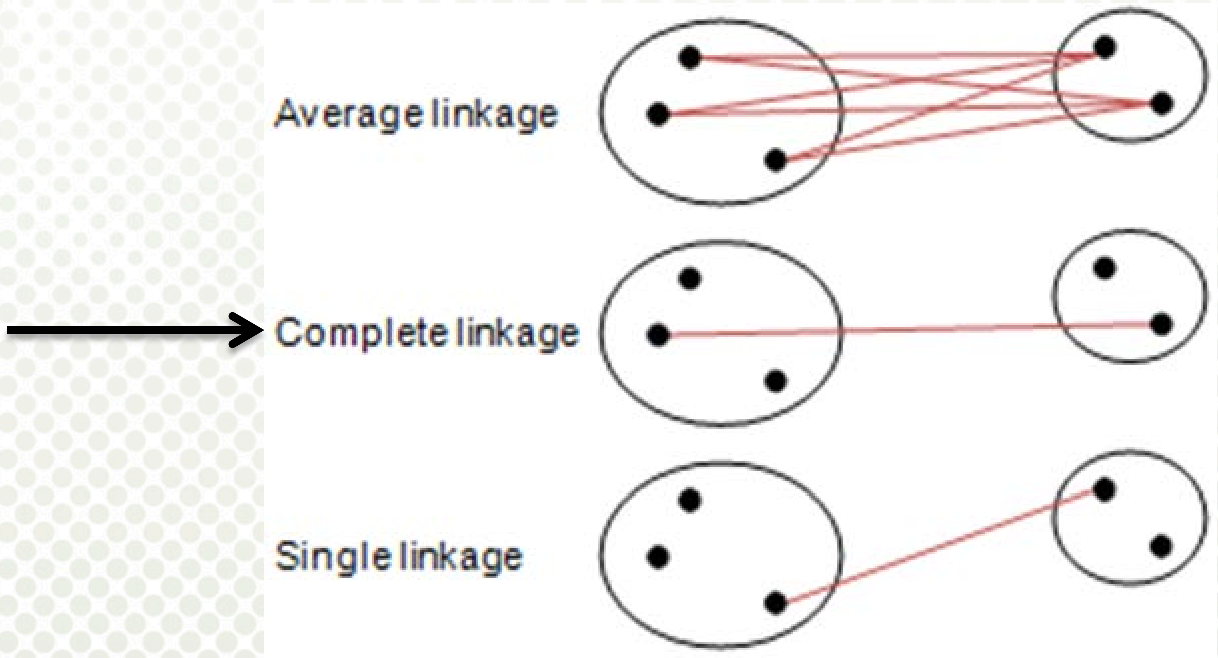
Figure 3.1 Distance given by the Pythagorean Theorem

Department of Animal and Aquaculture Sciences
UNIVERSITETET FOR MILJØ- OG BIOVITENSKAP

Hierarchical Clustering

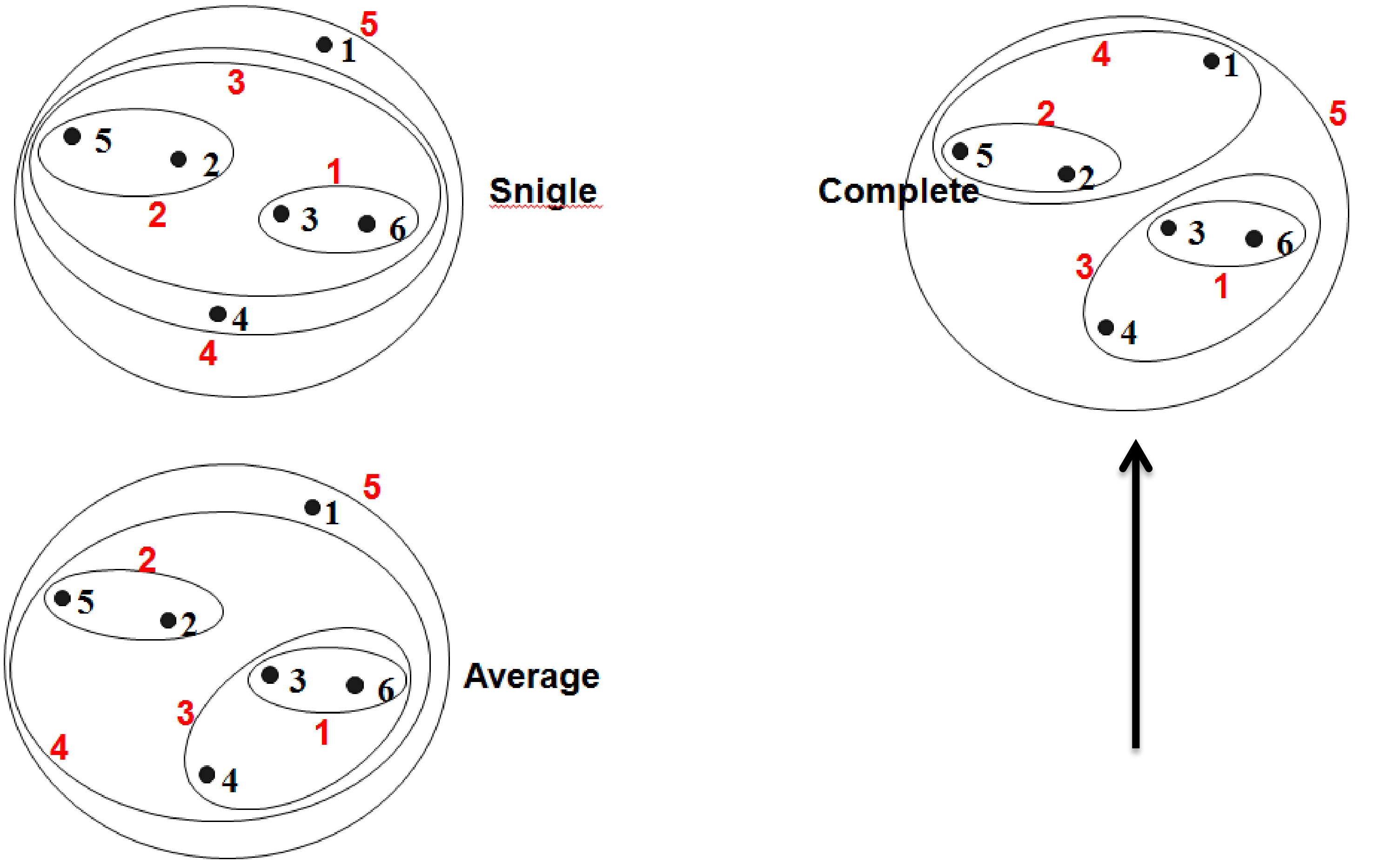
- Given training data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$

- Define a distance metric between points in inputs space. Common measures is Euclidean distance, either by



Department of Animal and Aquaculture Sciences
UNIVERSITETET FOR MILJØ- OG BIOVITENSKAP

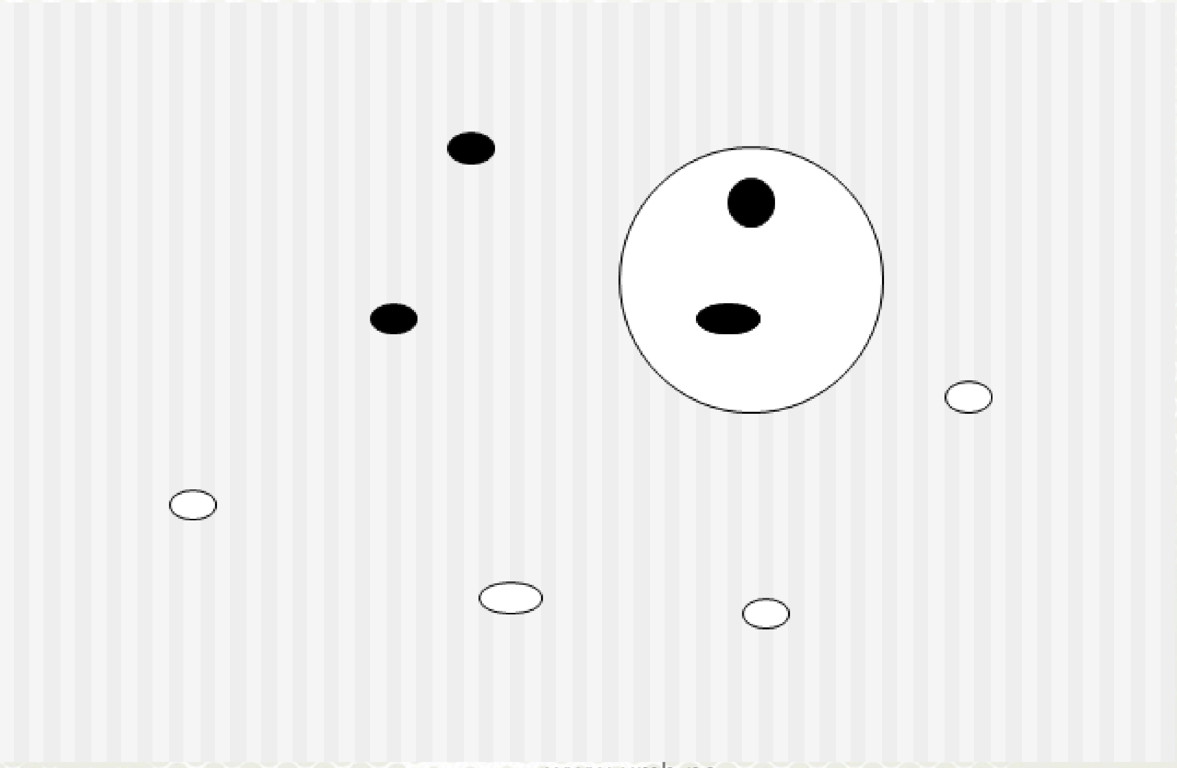
Hierarchical Clustering: Comparison

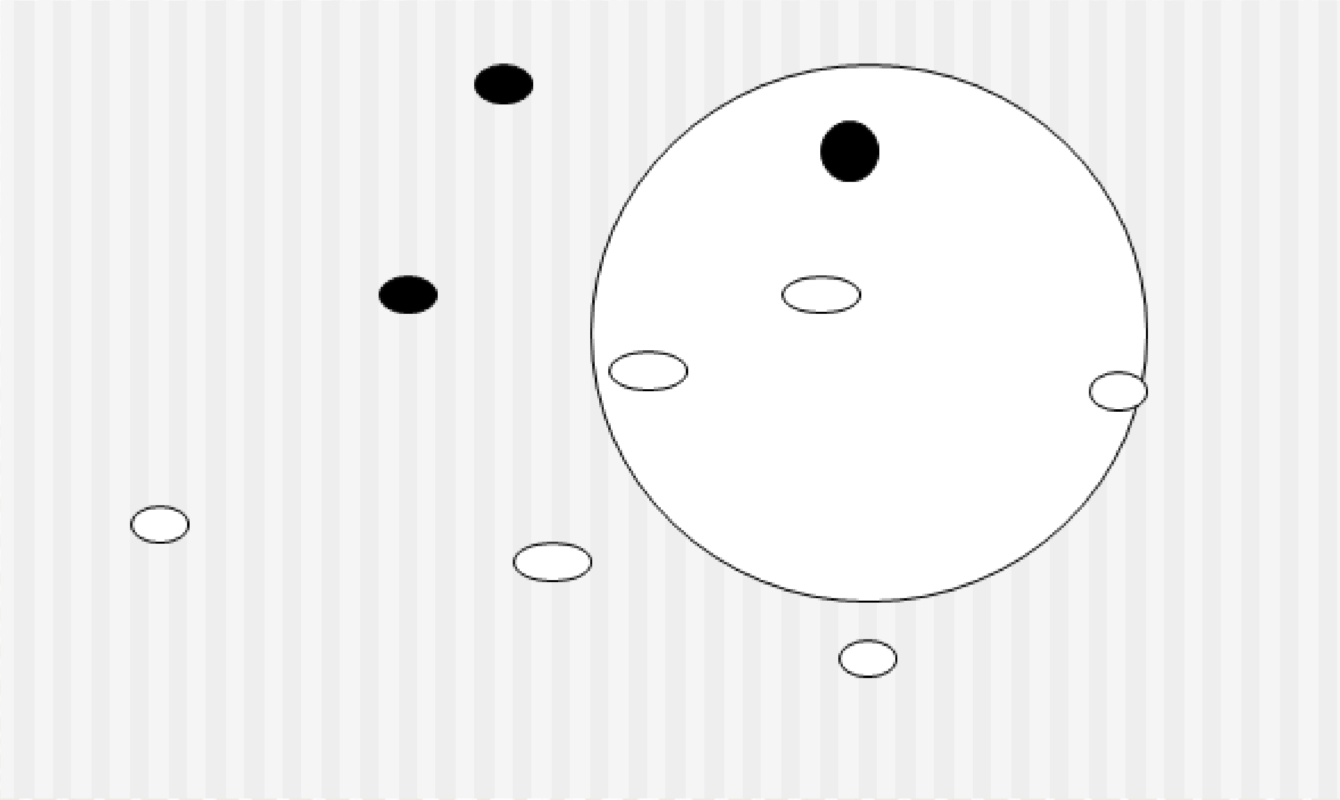


Department of Animal and Aquaculture Sciences
UNIVERSITETET FOR MILJØ- OG BIOVITENSKAP

Hierarchical Clustering

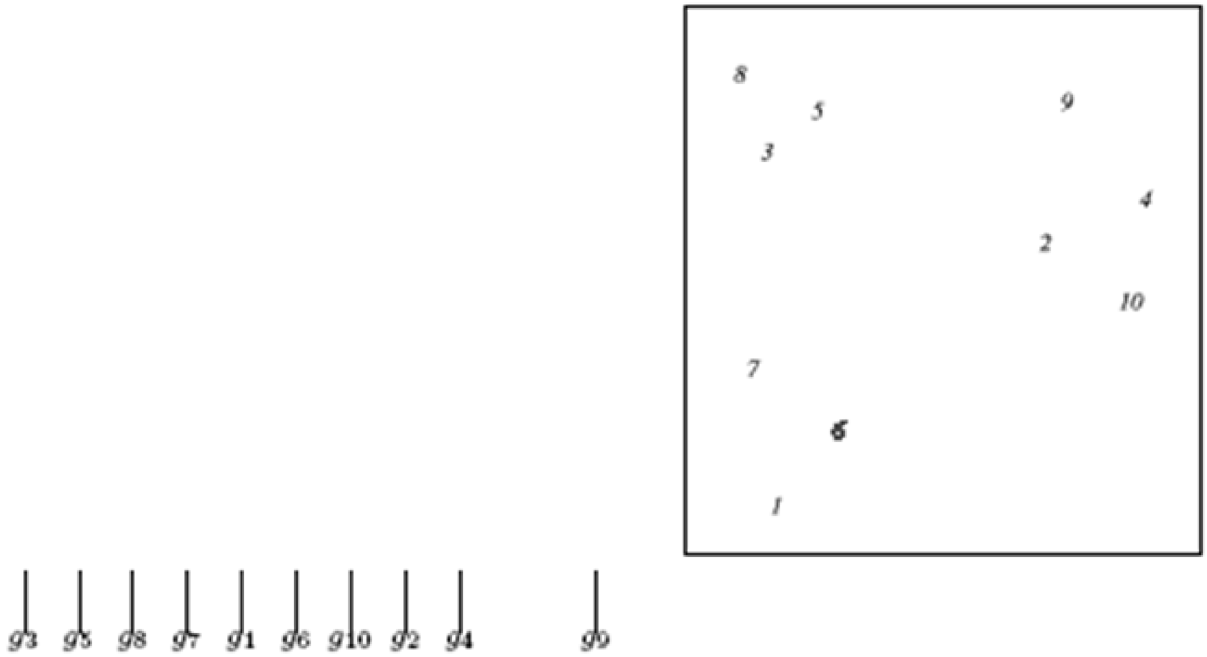
- Given test point $\mathbf{x}$
- Find the K nearest training inputs $\mathbf{x}_1, \ldots, \mathbf{x}_N$ to $\mathbf{x}$ given the distance metric $D(\mathbf{x}, \mathbf{x}_i)$



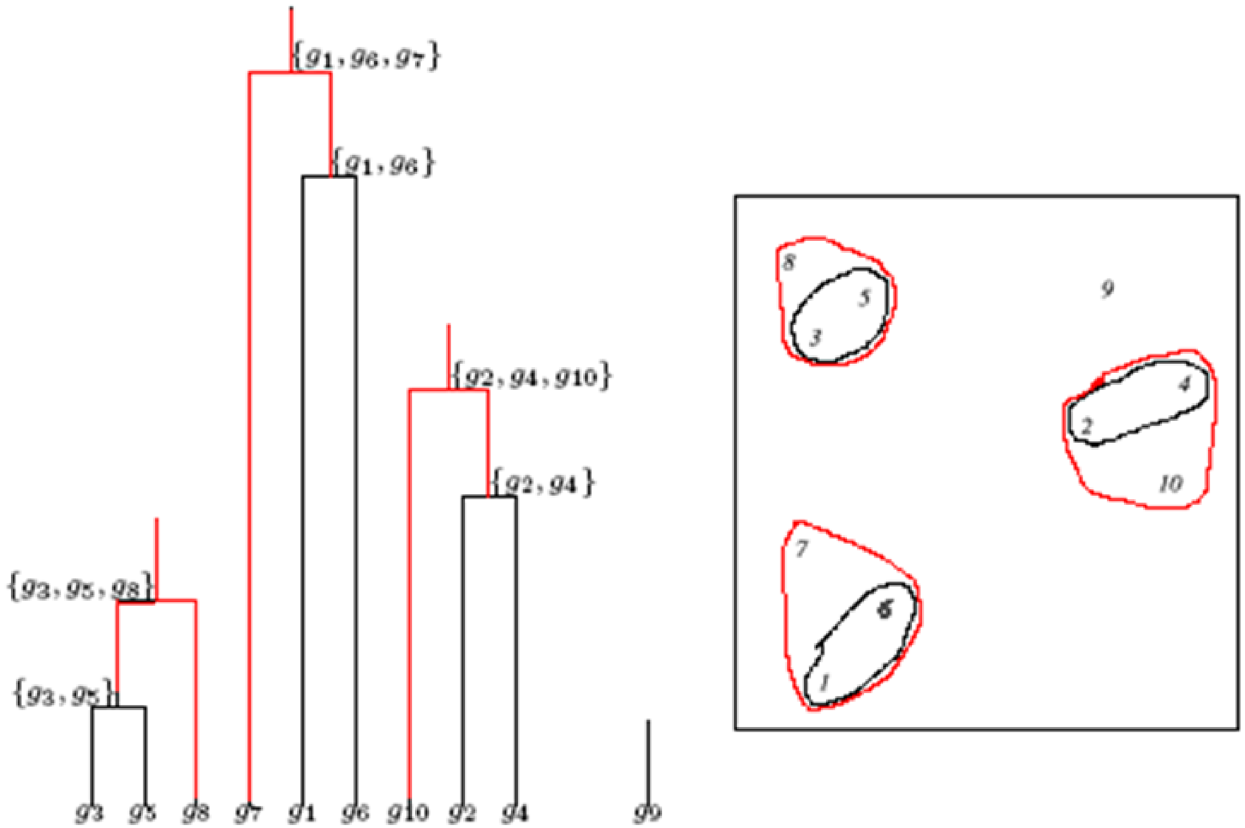


UNIVERSITETET FOR MILJØ- OG BIOVITENSKAP
Department of Animal and Aquaculture Sciences

Hierarchical Clustering: Example

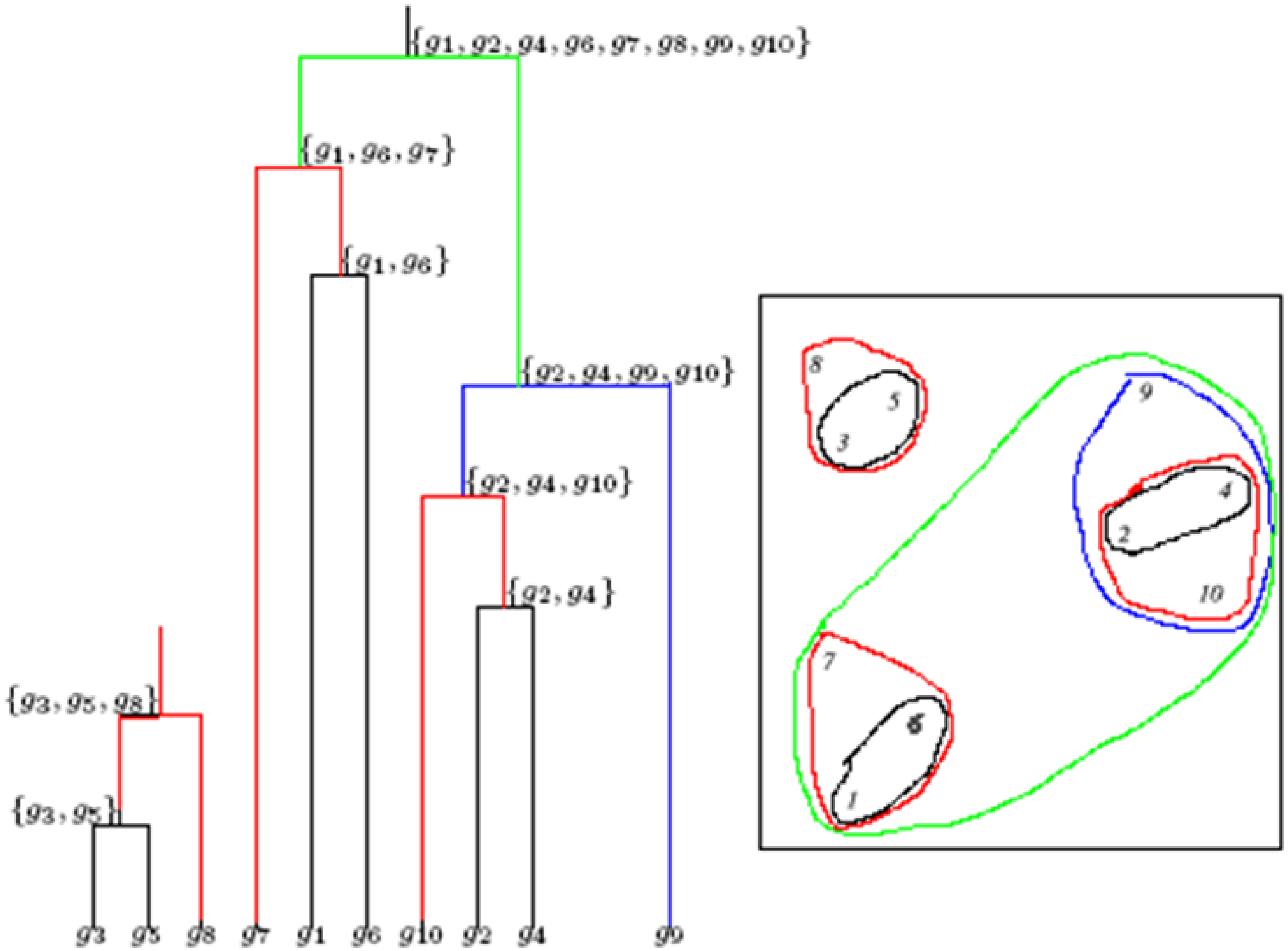


Hierarchical Clustering: Example



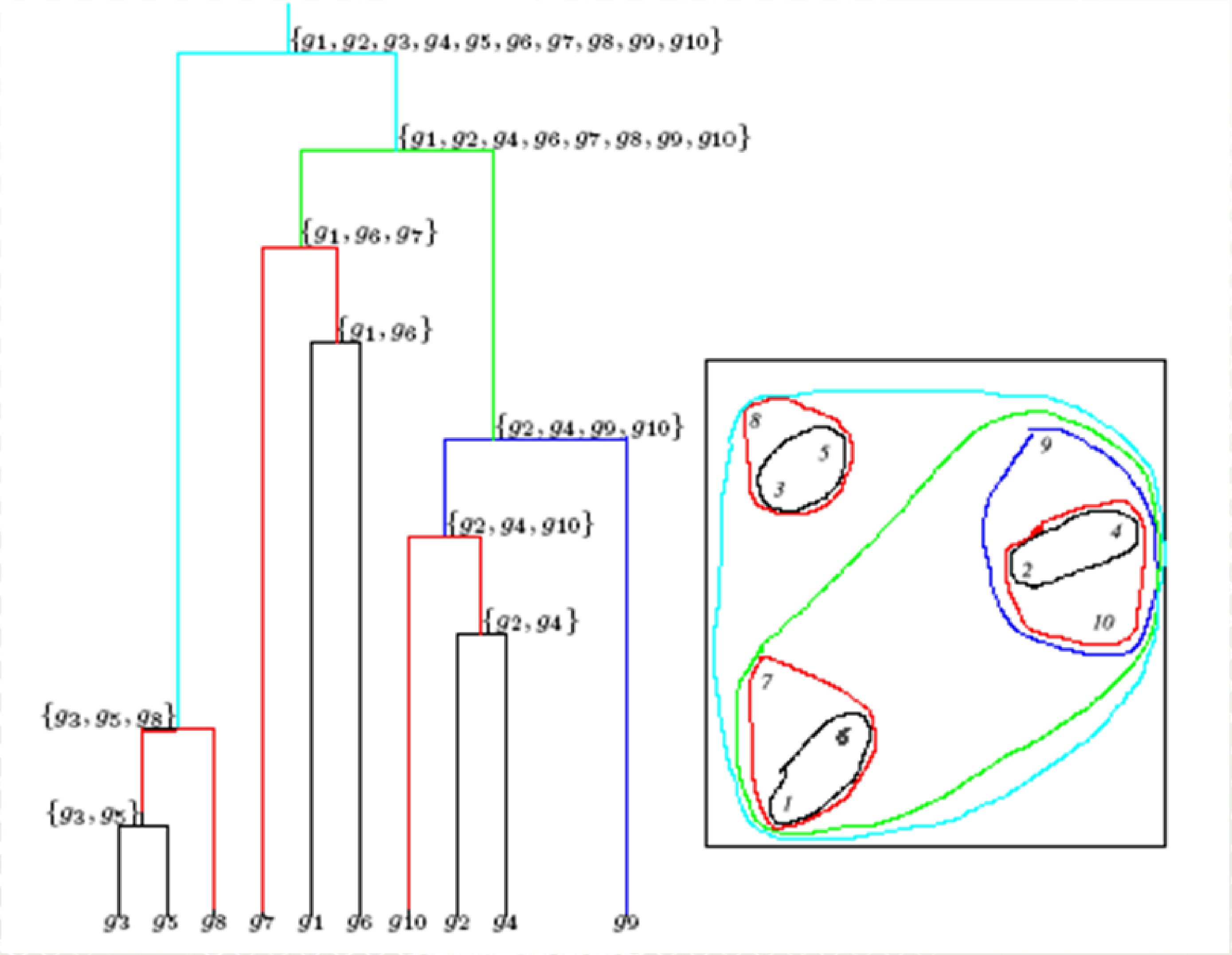
Department of Animal and Aquaculture Sciences
UNIVERSITETET FOR MILJØ- OG BIOVITENSKAP

Hierarchical Clustering: Example



Department of Animal and Aquaculture Sciences
UNIVERSITETET FOR MILJØ- OG BIOVITENSKAP

Hierarchical Clustering: Example



K-means vs hierarchical clustering

- This method differs from the hierarchical clustering in many ways. In particular,
- There is no hierarchy, the data are partitioned. You will be presented only with the final cluster membership for each case.
- There is no role for the dendrogram in k-means clustering.
- You must supply the number of clusters (k) into which the data are to be grouped.

Department of Animal and Aquaculture Sciences
UNIVERSITETET FOR MILJØ- OG BIOVITENSKAP

Implementation of Clustering in SNP imputation

Hap.	SNP1	SNP2	….	SNPn	Cluster
1	2	1	….	1	?
2	1	1	….	2	?
3	2	2	….	2	?
4	1	1	….	1	?
5	1	1	….	2	?

- Note that, the missing SNPs (missing-ness) can also used as a predictor in a clustering analysis

Implementation of Clustering in SNP imputation

- Example 3.2 (Clustering using complete linkage and the Euclidean distance)

- e.g correlation distance between haplotype 1 and haplotype 3:
- $d\,(h_1,h_3)\ = [1-\ cor(h_1,h_3)] * 10$

- $D = \{d_{ik}\} = \begin{array}{c} hap_1 \\ hap_2 \\ hap_3 \\ hap_4 \\ hap_5 \end{array} \begin{bmatrix} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 10 & 10 & (2) & 8 & 0 \end{bmatrix}$

Department of Animal and Aquaculture Sciences
UNIVERSITETET FOR MILJØ- OG BIOVITENSKAP

Clustering in SNP imputation

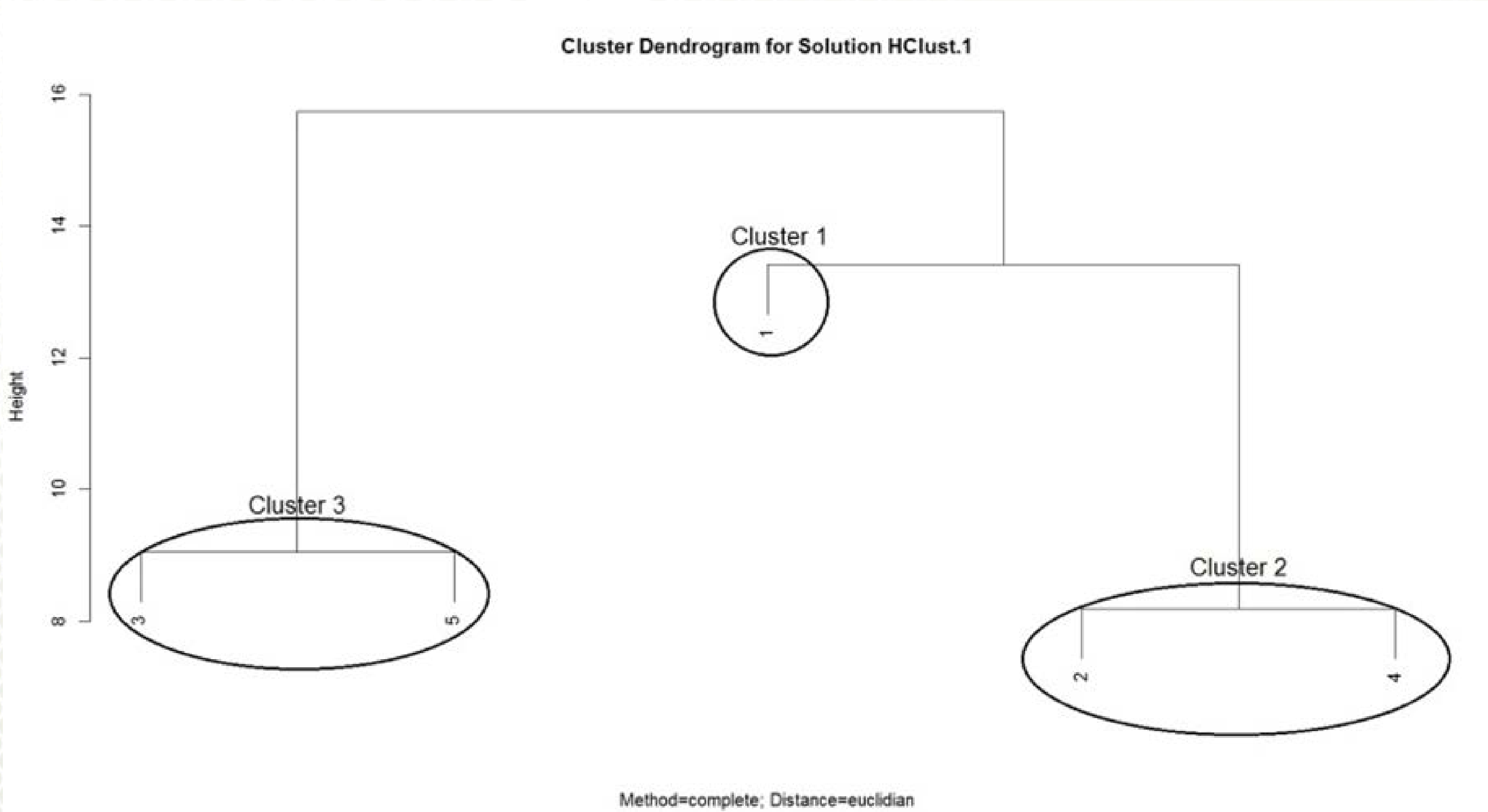
- $d_{(35)1} = \max\{d_{31}, d_{51}\} = \max\{3{,}10\} = 10$
- $d_{(35)2} = \max\{d_{32}, d_{52}\} = 10$
- $d_{(35)4} = \max\{d_{34}, d_{54}\} = 9$
- $$\begin{array}{c} hap_{(35)} \\ hap_1 \\ hap_2 \\ hap_4 \end{array} \begin{bmatrix} 0 & & & \\ 10 & 0 & & \\ 10 & 9 & 0 & \\ 9 & 6 & (5) & 0 \end{bmatrix} \dashrightarrow \begin{array}{c} (35) \\ (24) \\ 1 \end{array} \begin{bmatrix} 0 & & \\ 10 & 0 & \\ 10 & 9 & 0 \end{bmatrix}$$

UNIVERSITETET FOR MILJØ- OG BIOVITENSKAP
Department of Animal and Aquaculture Sciences

R commands

- hclust(dist(model.matrix(~-1 + hap1+hap2+hap3+hap4+hap5,Dataset)), method= "complete")
- plot(HClust.1, main= "Cluster Dendrogram for Solution HClust.1", sub="Method=complete;Distance=**euclidian**")
- Dataset$hclus.label <- assignCluster(model.matrix(~-1 + hap1 + hap2 + hap3 + hap4 + hap5, Dataset), Dataset, cutree(HClust.1, k = 3))

Clustering in SNP imputation



UNIVERSITETET FOR MILJØ- OG BIOVITENSKAP
Department of Animal and Aquaculture Sciences

Clustering in SNP imputation

Hap.	SNP1	SNP2	….	SNPn	Cluster
1	2	1	….	1	1
2	1	1	….	2	2
3	2	2	….	2	3
4	1	1	….	1	2
5	1	1	….	2	3

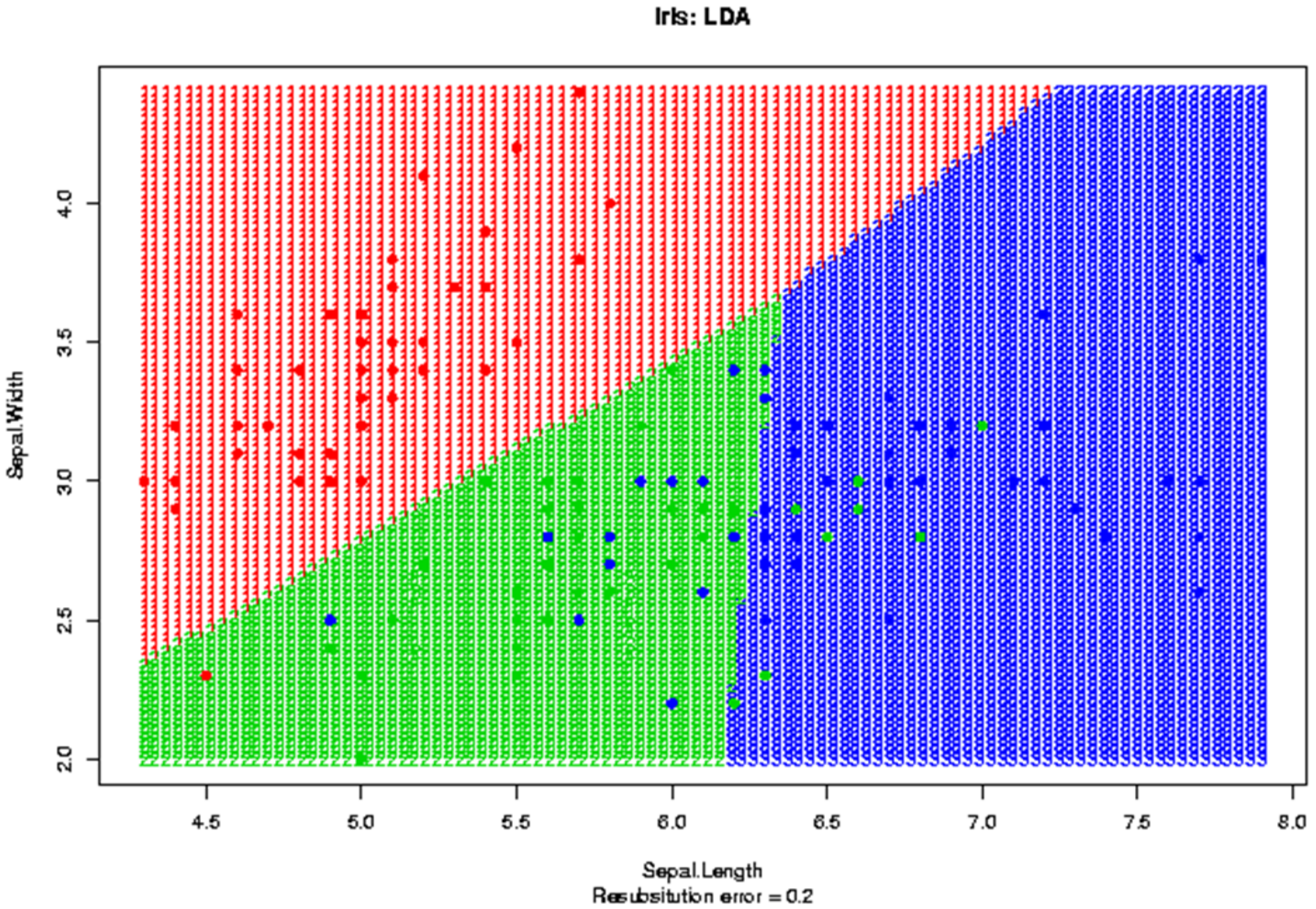
Department of Animal and Aquaculture Sciences
UNIVERSITETET FOR MILJØ- OG BIOVITENSKAP

Comparison between the methods LDA and Clustering

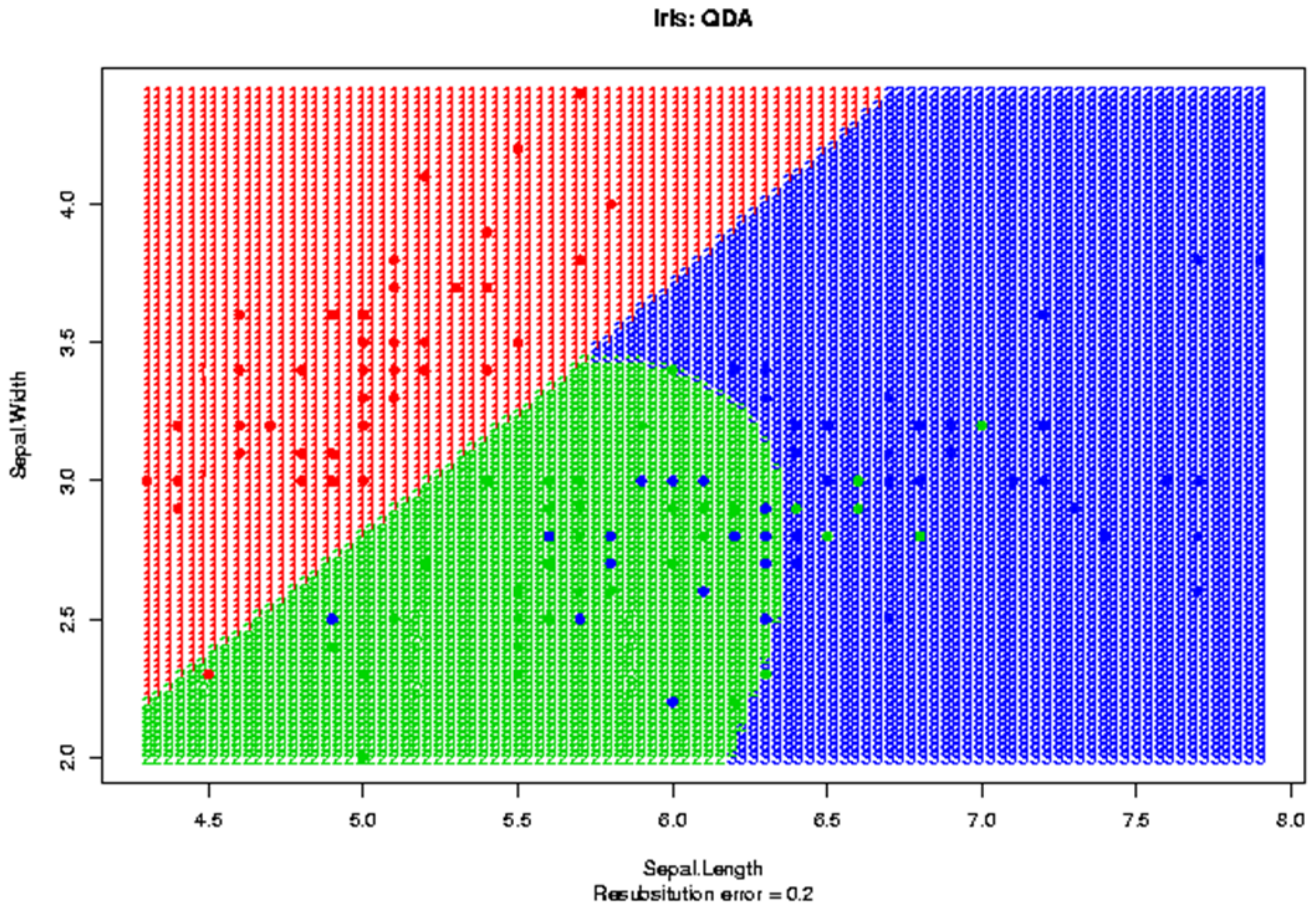
- Ex. Iris Data
  - Y: 3 species,
    - Iris setosa (red), versicolor (green), and virginica (blue).
  - X: 2 variables
    - Sepal length and width

Department of Animal and Aquaculture Sciences
UNIVERSITETET FOR MILJØ- OG BIOVITENSKAP

Example: Linear discriminant analysis



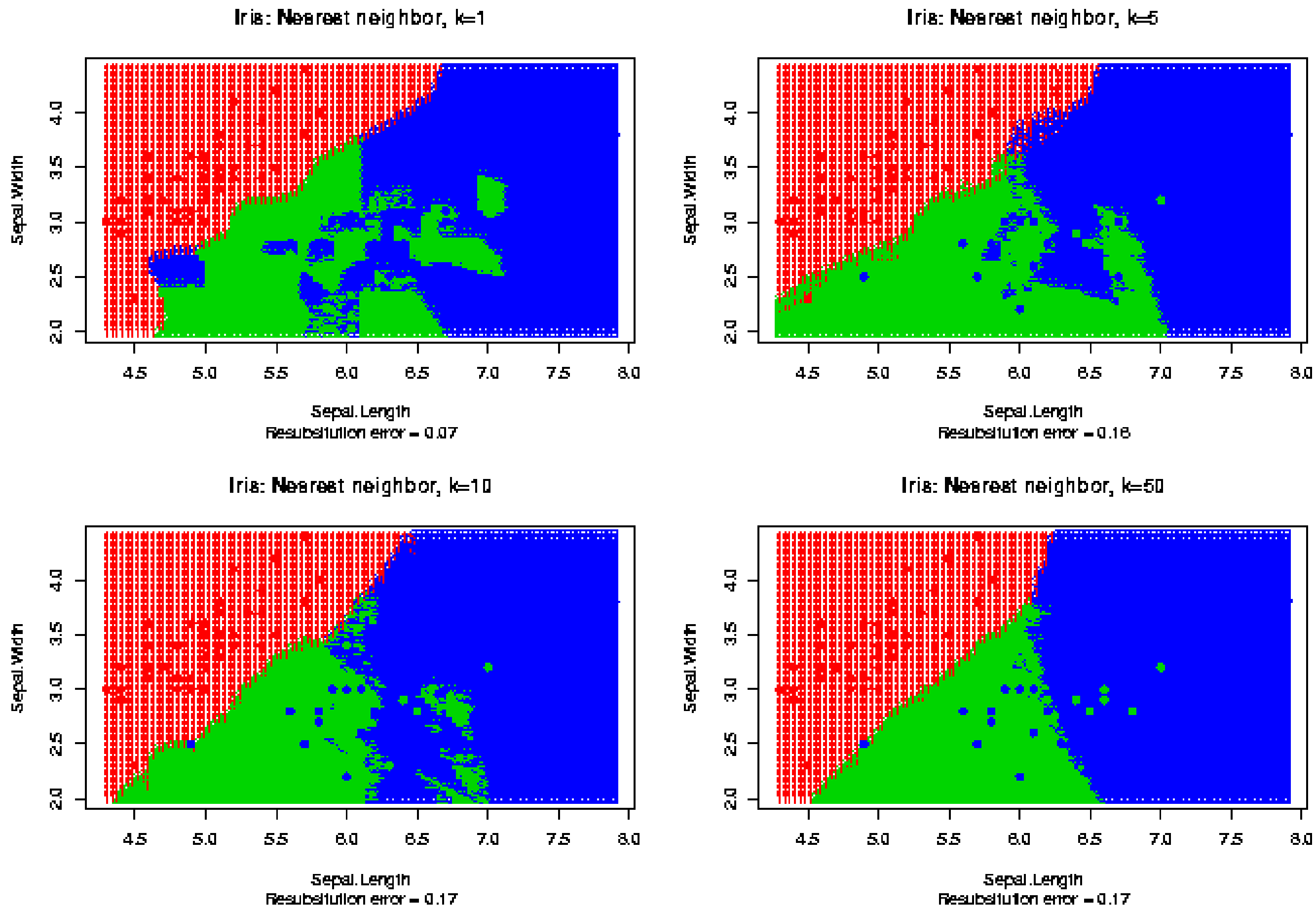
Example: Quadratic discriminant analysis



Department of Animal and Aquaculture Sciences
UNIVERSITETET FOR MILJØ- OG BIOVITENSKAP

Department of Animal and Aquaculture Sciences
UNIVERSITETET FOR MILJØ- OG BIOVITENSKAP

Example: Nearest neighbor classifier

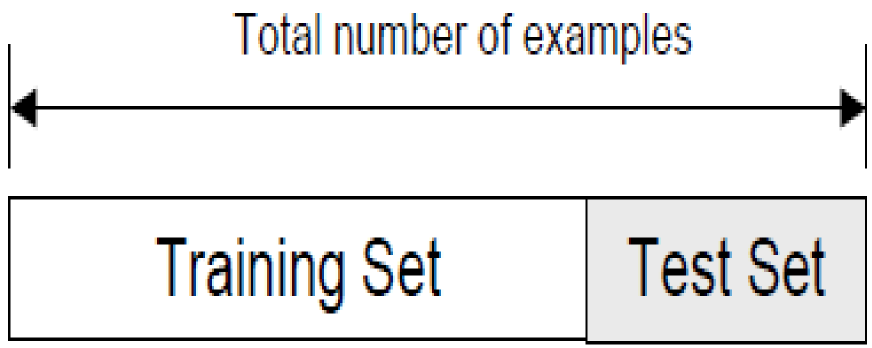


Overview

- Introduction
  - Imputation and Multiple imputation
  - Genotype imputation
  - Aim of the study
- Materials and Simulations
- Methods
  - Linear discriminant analysis
  - Clustering analysis
- Validation
- Results
- Discussion and Conclusion
- Questions

Validation

- **The holdout method**
- Usually we using 50% training data-set in this study, except in the last experiment where we measuring the effect of the size of the training dataset (where we varying the size of the training dataset (n)).



Department of Animal and Aquaculture Sciences
UNIVERSITETET FOR MILJØ- OG BIOVITENSKAP

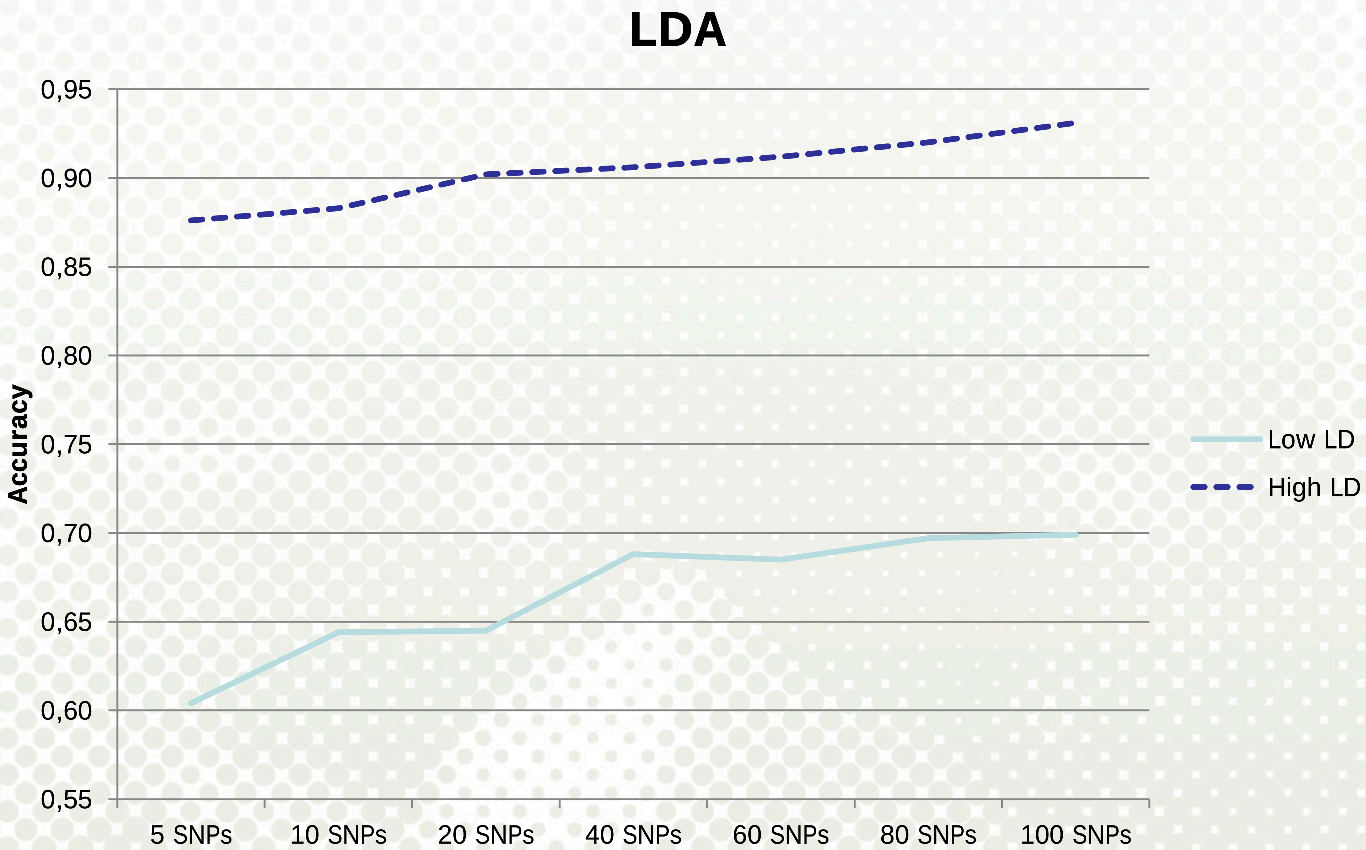
Overview

- Introduction
  - Imputation and Multiple imputation
  - Genotype imputation
  - Aim of the study
- Materials and Simulations
- Methods
  - Linear discriminant analysis
  - Clustering analysis
- Validation
- Results
- Discussion and Conclusion
- Question

Department of Animal and Aquaculture Sciences
UNIVERSITETET FOR MILJØ- OG BIOVITENSKAP

4.0 RESULTS

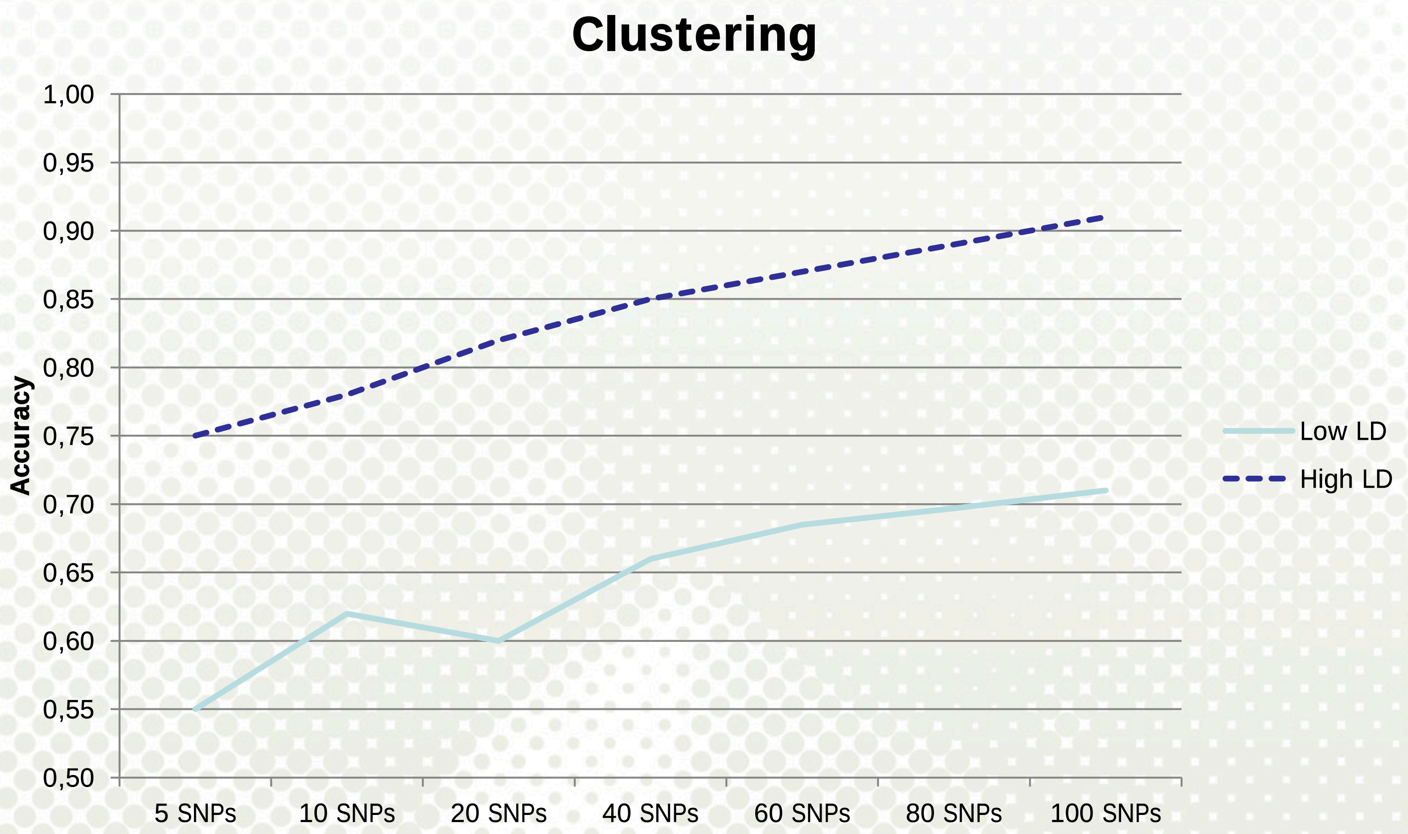
- **Figure 1:** shows the effects of size of haplotype block (number of SNPs per haplotype), on imputation accuracy rate (AR) using low and high linkage disequilibrium dataset (LLD, HLD).



Department of Animal and Aquaculture Sciences
UNIVERSITETET FOR MILJØ- OG BIOVITENSKAP

4.0 RESULTS

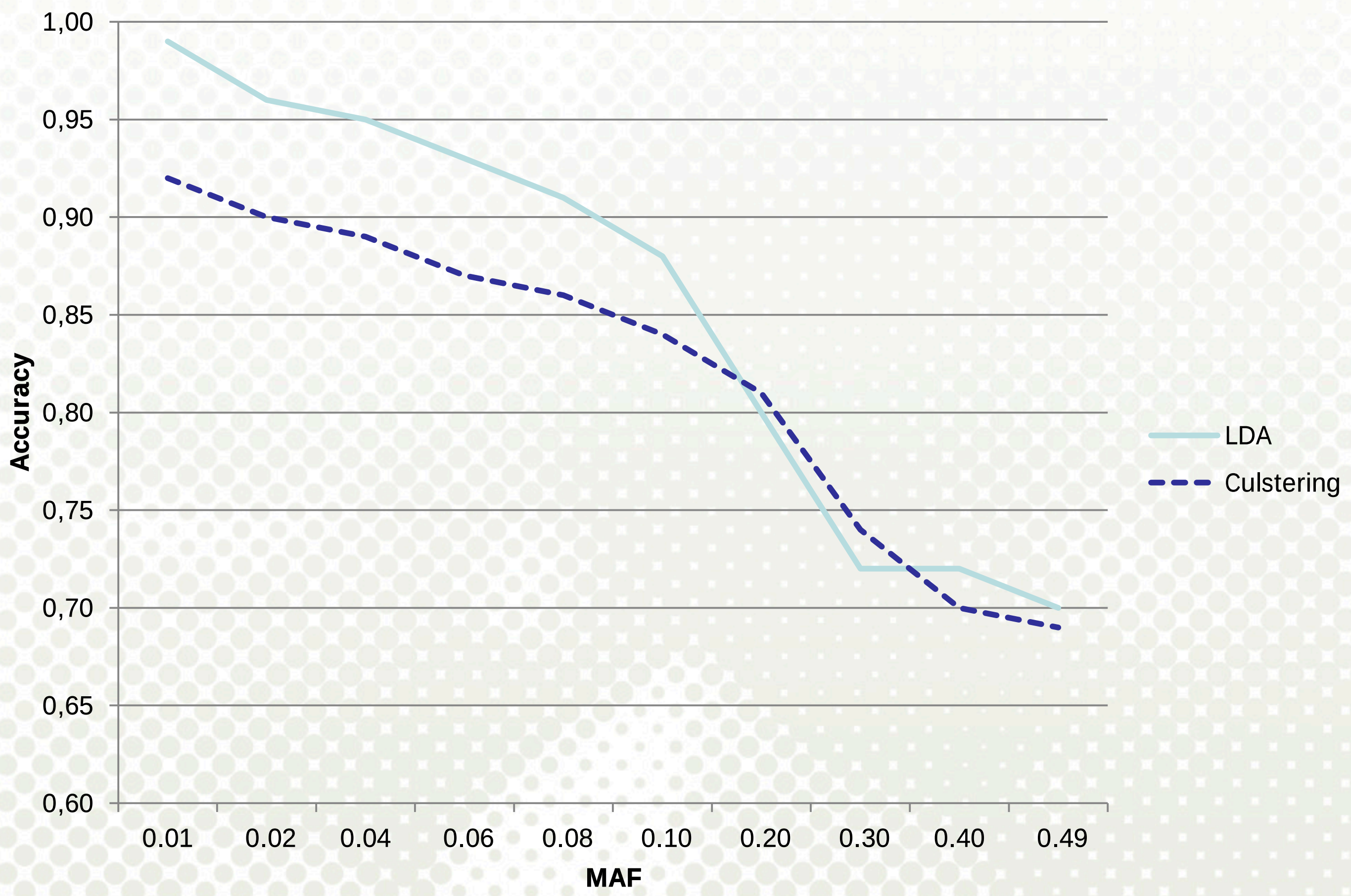
- **Figure 2:** Shows the effects of number of SNPs surrounding the missing one, in imputation accuracy rate (AR) using low and high linkage disequilibrium dataset (LLD, HLD).



Department of Animal and Aquaculture Sciences
UNIVERSITETET FOR MILJØ- OG BIOVITENSKAP

4.0 RESULTS

- **The effects of Minor allele frequency (MAF): Figure 3.**



4.0 RESULTS

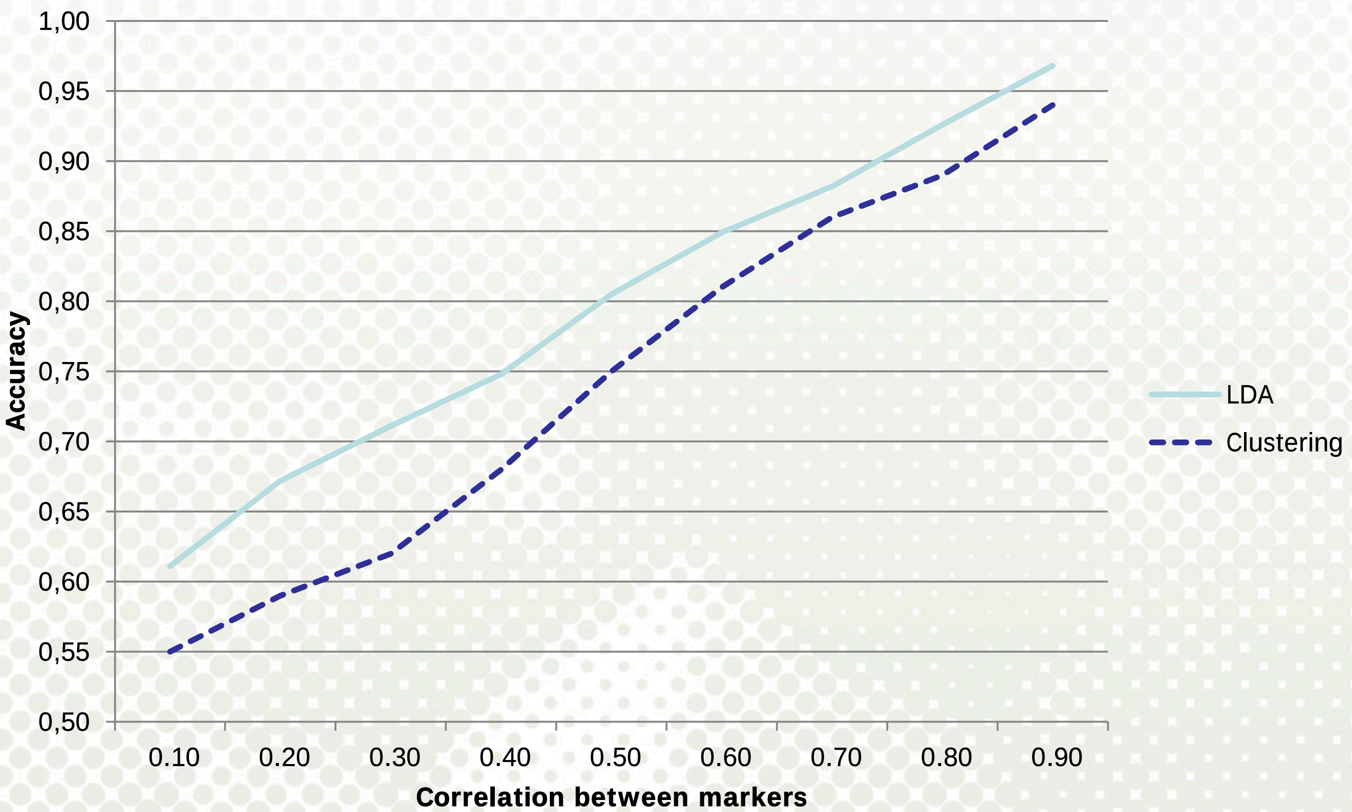
MAF effect

- It seems that AR is much more accurate when MAF is low compared to when it is high. A lower MAF usually corresponds to a stronger LD with nearby markers and the recombination plays a primary role in LD decay (Yu-Fang Pei., 2008).

Department of Animal and Aquaculture Sciences
UNIVERSITETET FOR MILJØ- OG BIOVITENSKAP

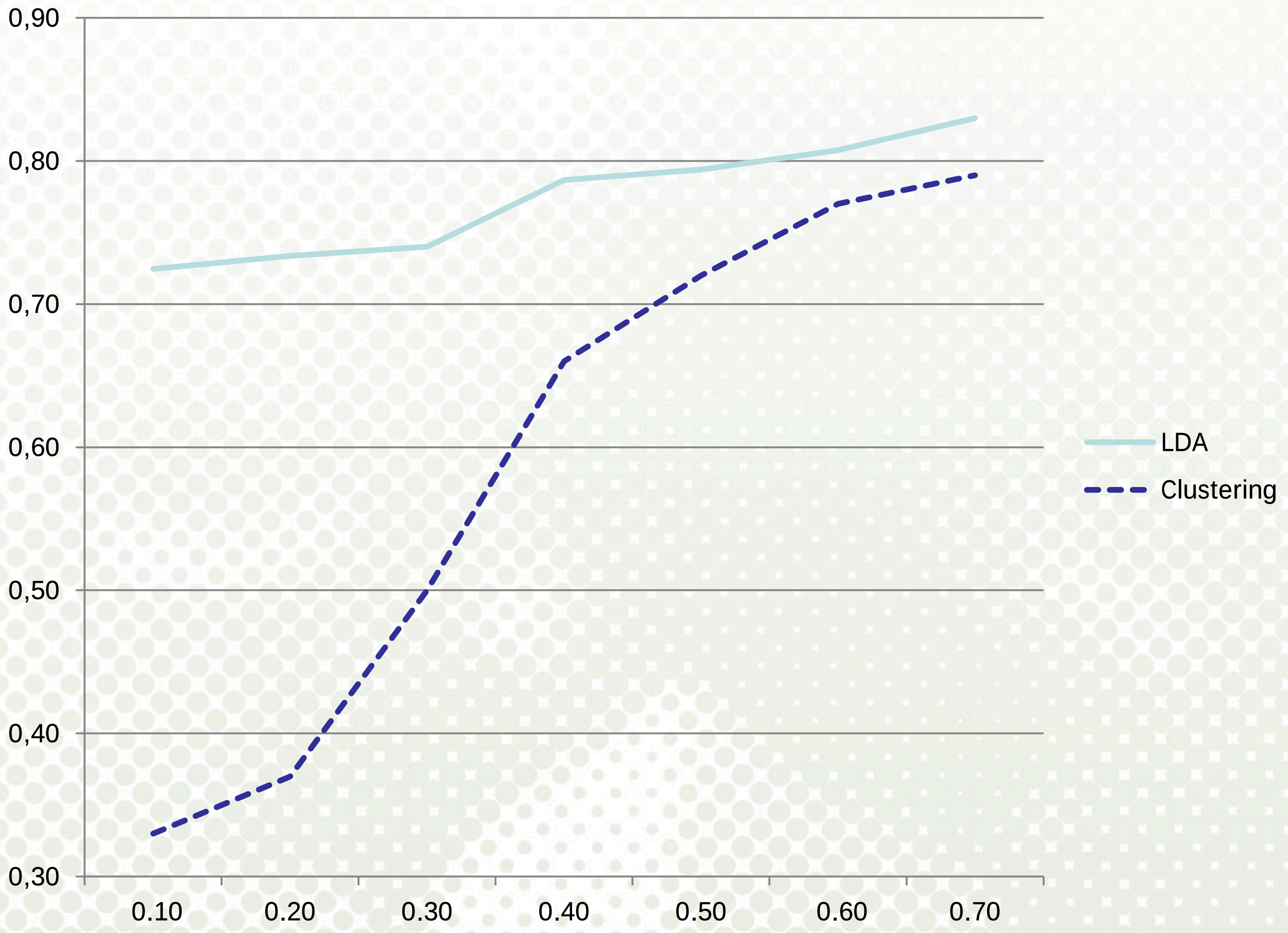
4.0 RESULTS

- **The effects of marker density (MD)**
- Here, we measure the effect of Marker density by varying the correlation between markers (SNPs).



4.0 RESULTS

- **The effects of reference sample size (n):** Figure 5.



Department of Animal and Aquaculture Sciences
UNIVERSITETET FOR MILJØ- OG BIOVITENSKAP

Overview

- Introduction
    - Imputation and Multiple imputation
    - Genotype imputation
    - Aim of the study
- Materials and Simulations
- Methods
    - Linear discriminant analysis
    - Clustering analysis
- Validation
- Results
- Discussion and Conclusion
- Questions

DISCUSSION AND CONCLUSION

- The performance of the elementary imputation methods, clustering and discrimination is generally good. However, to compare the performance of each algorithm with the currently used methods like in MACH, BEAGLE and IMPUTE, more test experiments are needed to be conducted.

- In low LD region, the clustering-based method can use the correlation between records instead of the correlation between markers in the imputation process.

- The Discriminant-based method also can handle numerical and categorical data simultaneously without rounding-up the results (which can affect the accuracy of imputation).

UNIVERSITETET FOR MILJØ- OG BIOVITENSKAP
Department of Animal and Aquaculture Sciences

DISCUSSION AND CONCLUSION

- In optimal state of genotype data (in High LD, low MAF, and high density haplotype blokes) both methods (Clustering and discrimination) were working efficiently, and the accuracy can reached 89 %.

- Results obtained had many similarities with those obtained both from Discriminant-based imputation and Clustering-based SNP imputation approaches in similar datasets.

DISCUSSION AND CONCLUSION

- Finally, searching for a new technique and a new application or a new demonstration of Discriminant and Clustering analysis was the main interest of this study because nowadays the application of the modern statistical techniques such LDA, Clustering, PCA, PLS... and etc., are so important considerations in the field of Bioinformatics and Applied statistic.

Thanks ☺

Contact

Medhat Mahmoud

Tel/ +49 38208 68 908

E-Mail/ mahmoud@fbn-dummerstorf.de

www.fbn-dummerstorf.de