UNISS
UNIVERSITÀ
DEGLI STUDI
DI SASSARI

DISCRIMINANT ANALYSIS

CORRADO DIMAURO, ALBERTO CESARANI

ATHENS, JANUARY 2025

In this lesson, and the next, we will talk about two multivariate statistical techniques:

The discriminant analysis (DA)

The cluster analysis (CA)

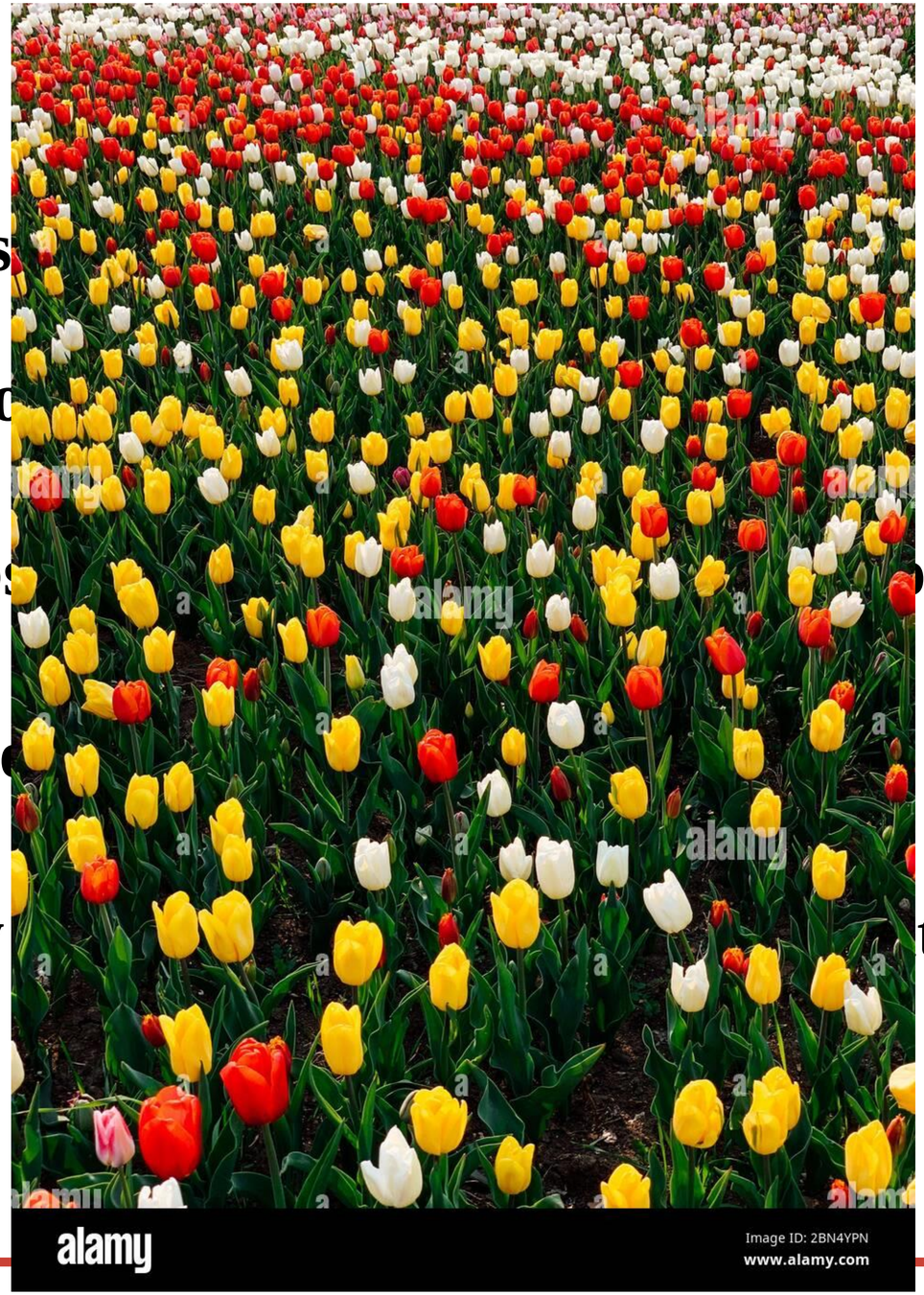
both techniques classify objects

Groups already exist

Groups do not exist

Corrado Dimauro, Alberto Cesarani
Università di Sassari

- **Classify observations us**
- **Obtain new classificatio**
- **Test the theory: are obs r not?**
- **Study the differences be**
- **Get an economical way to the groups they belong to**



Corrado Dimauro, Alberto Cesarani
Università di Sassari

EXAMPLES

versicolor

virginica

setosa

5

Let’s suppose we have a matrix of data **X** with p-variables measured on *n* experimental units

$$X = \begin{pmatrix} x_{11} & \cdot & \cdot & x_{1p} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ x_{n1} & \cdot & \cdot & x_{np} \end{pmatrix}$$

Cow	HG	HD	HLL	HC
1	128	133	134	126
2	130	135	139	126
3	131	133	134	125
4	131	135	137	133
5	129	133	133	130
6	141	143	144	138
7	132	134	137	127

Dim = ***nxp***

From X we extract the variance covariance matrix

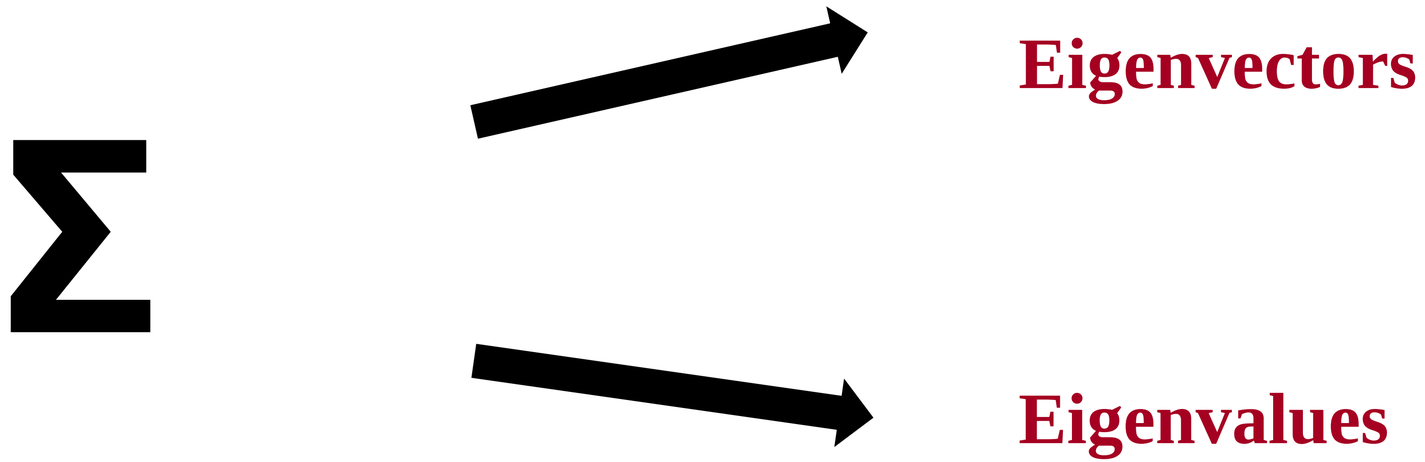
$$X = \begin{pmatrix} x_{11} & . & . & x_{1p} \\ . & . & . & . \\ . & . & . & . \\ x_{n1} & . & . & x_{np} \end{pmatrix}$$

Dim = ***n*** **x** ***p***

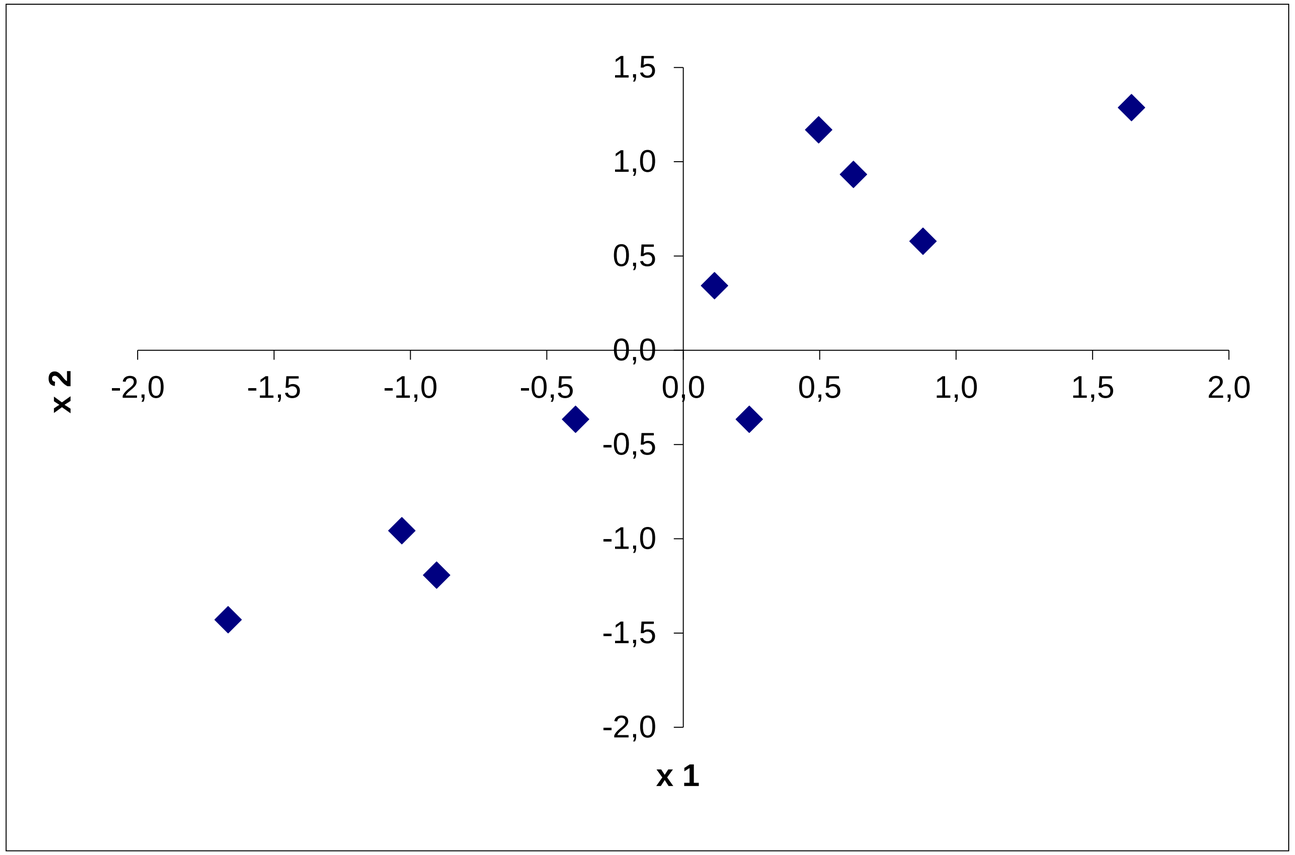
$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & . & \sigma_{1p} \\ \sigma_{12} & \sigma_{22} & . & \sigma_{2p} \\ . & . & . & . \\ \sigma_{1p} & \sigma_{2p} & . & \sigma_{pp} \end{pmatrix}$$

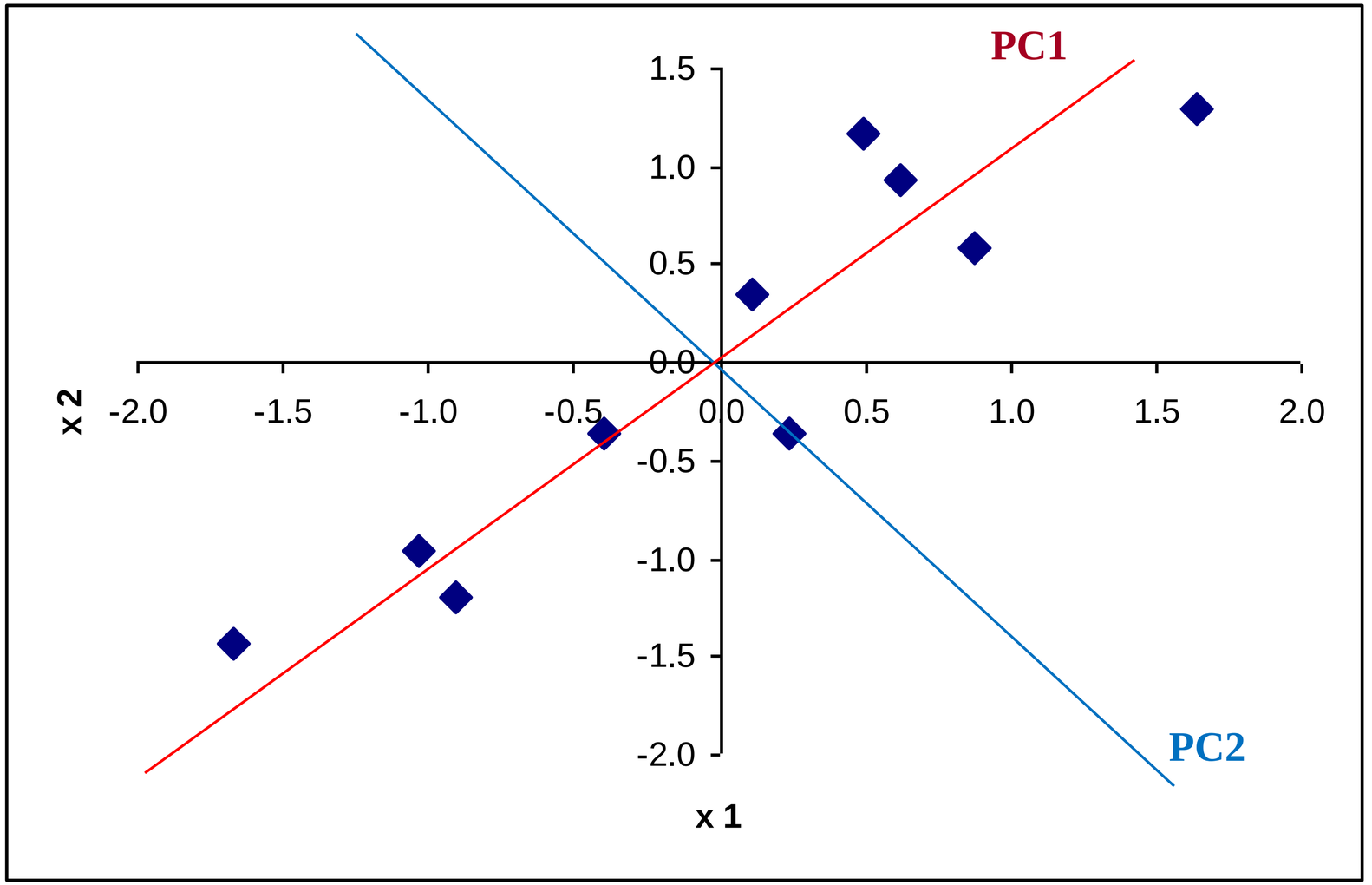
Dim = ***p*** **x** ***p***

From the variance covariance matrix

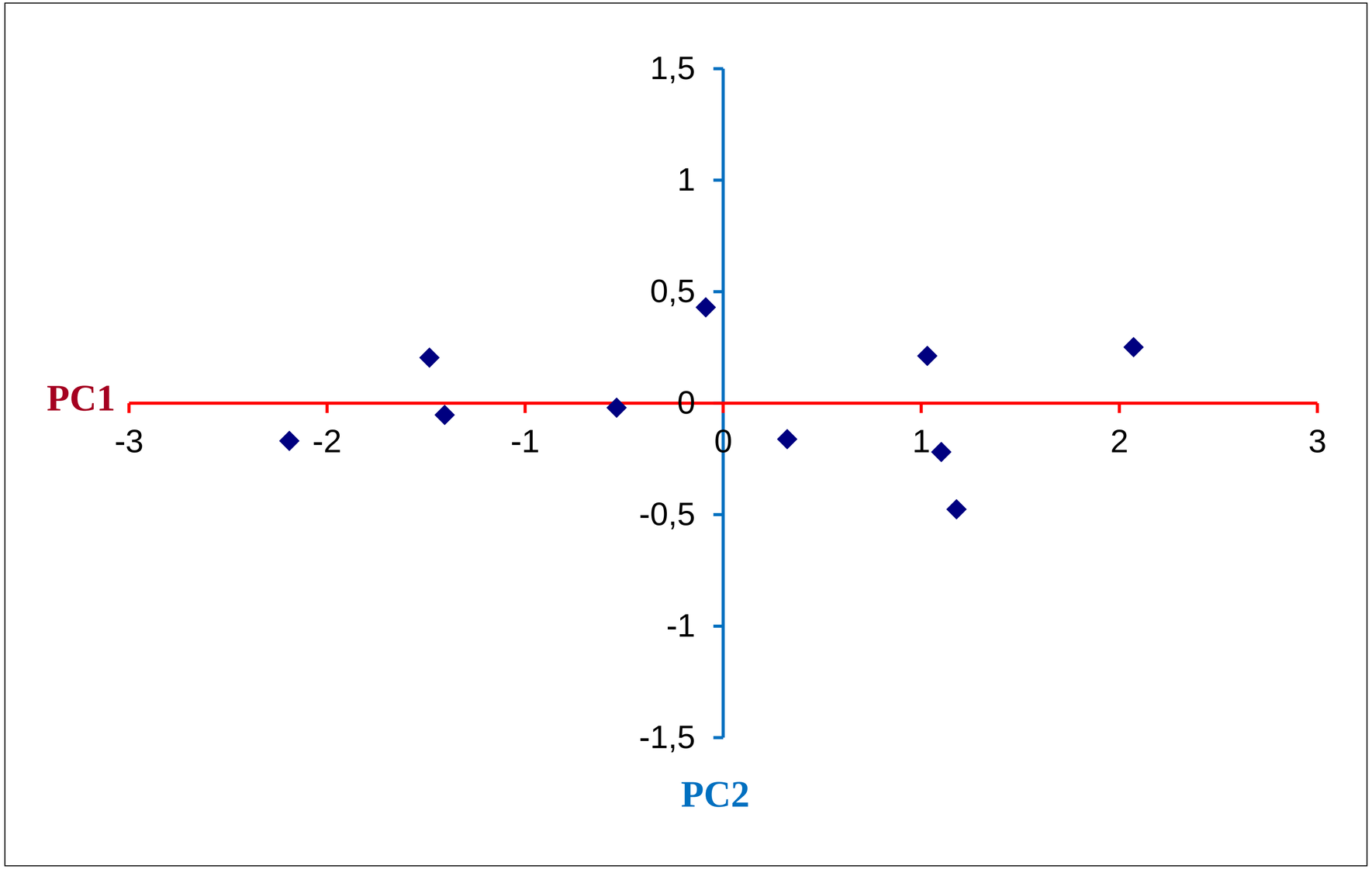


x1	x2
0,9	0,6
-1,7	-1,4
0,5	1,2
0,1	0,3
1,6	1,3
0,6	0,9
0,2	-0,4
-1,0	-1,0
-0,4	-0,4
-0,9	-1,2

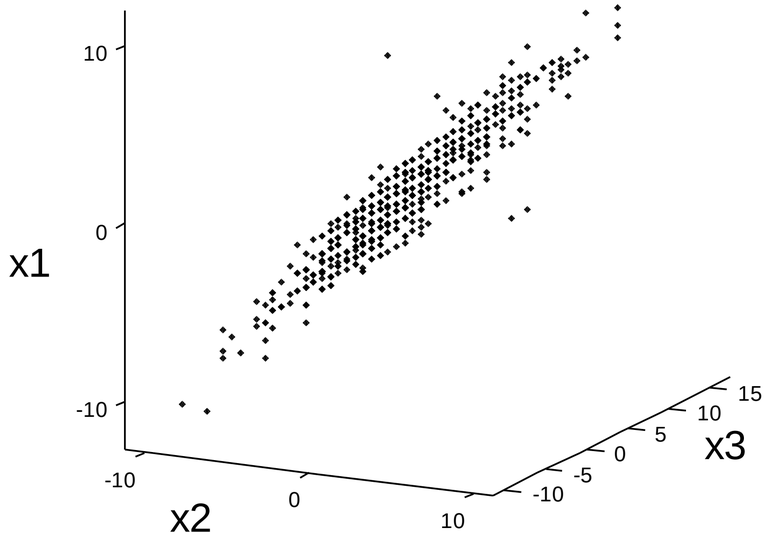


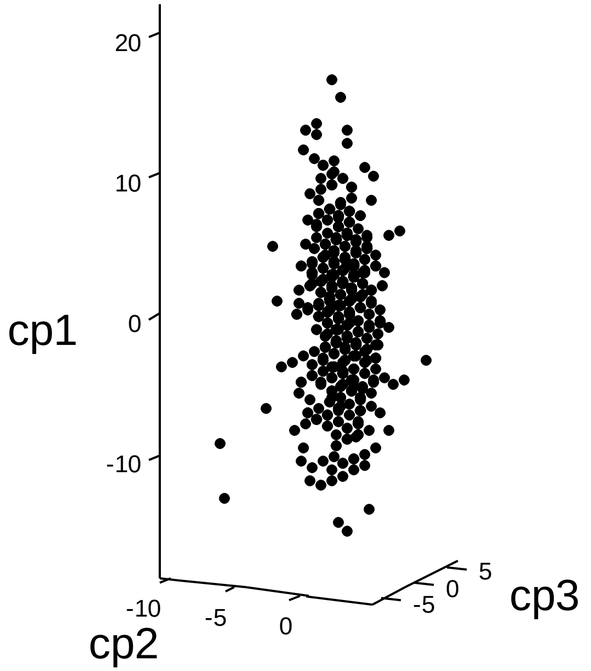

PC1
PC2
x 2
x 1
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0

After rotation



In three dimensions





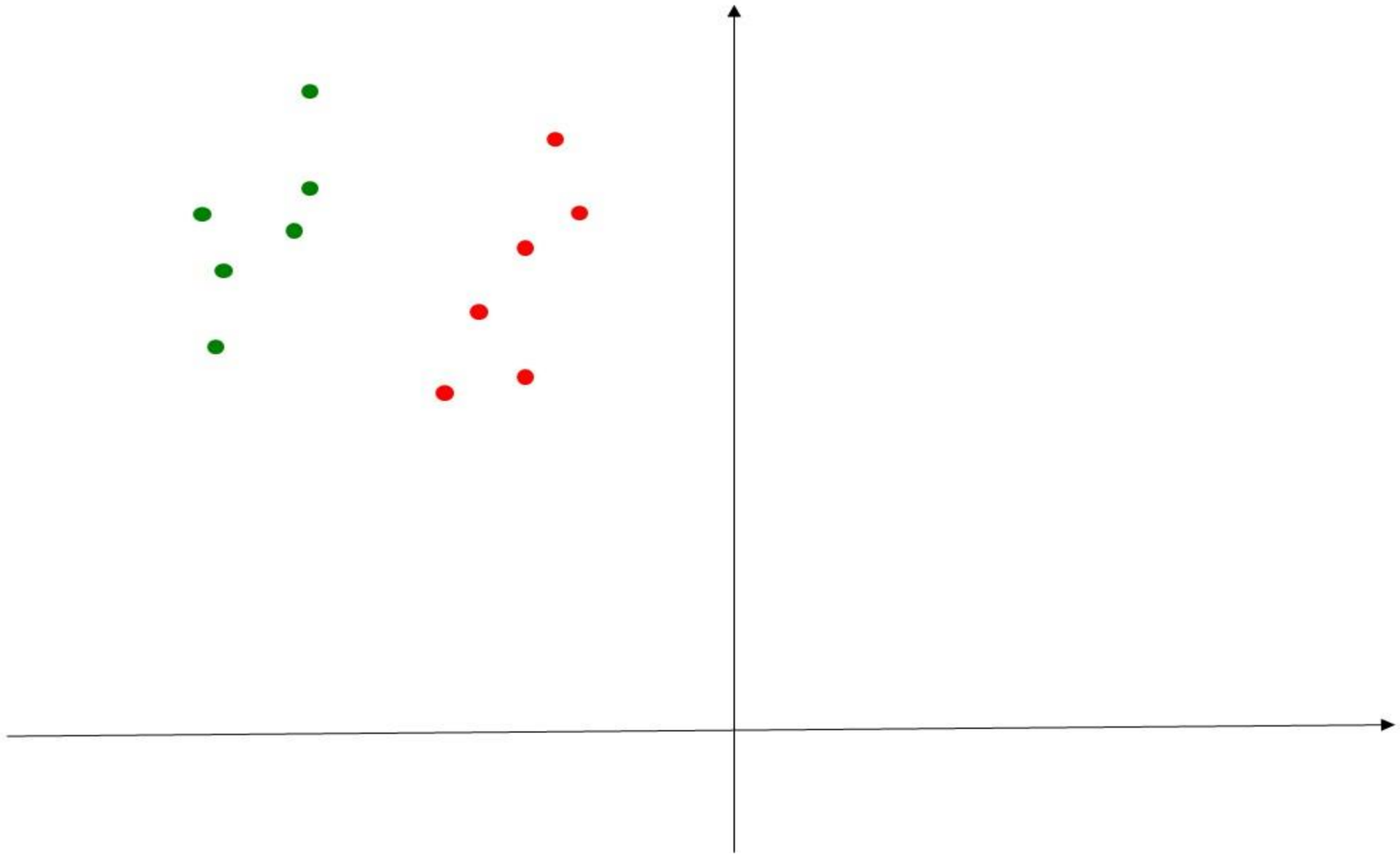
The GOAL of PCA is the rotation of the Cartesian reference system along the directions of maximum variability AMONG DATA

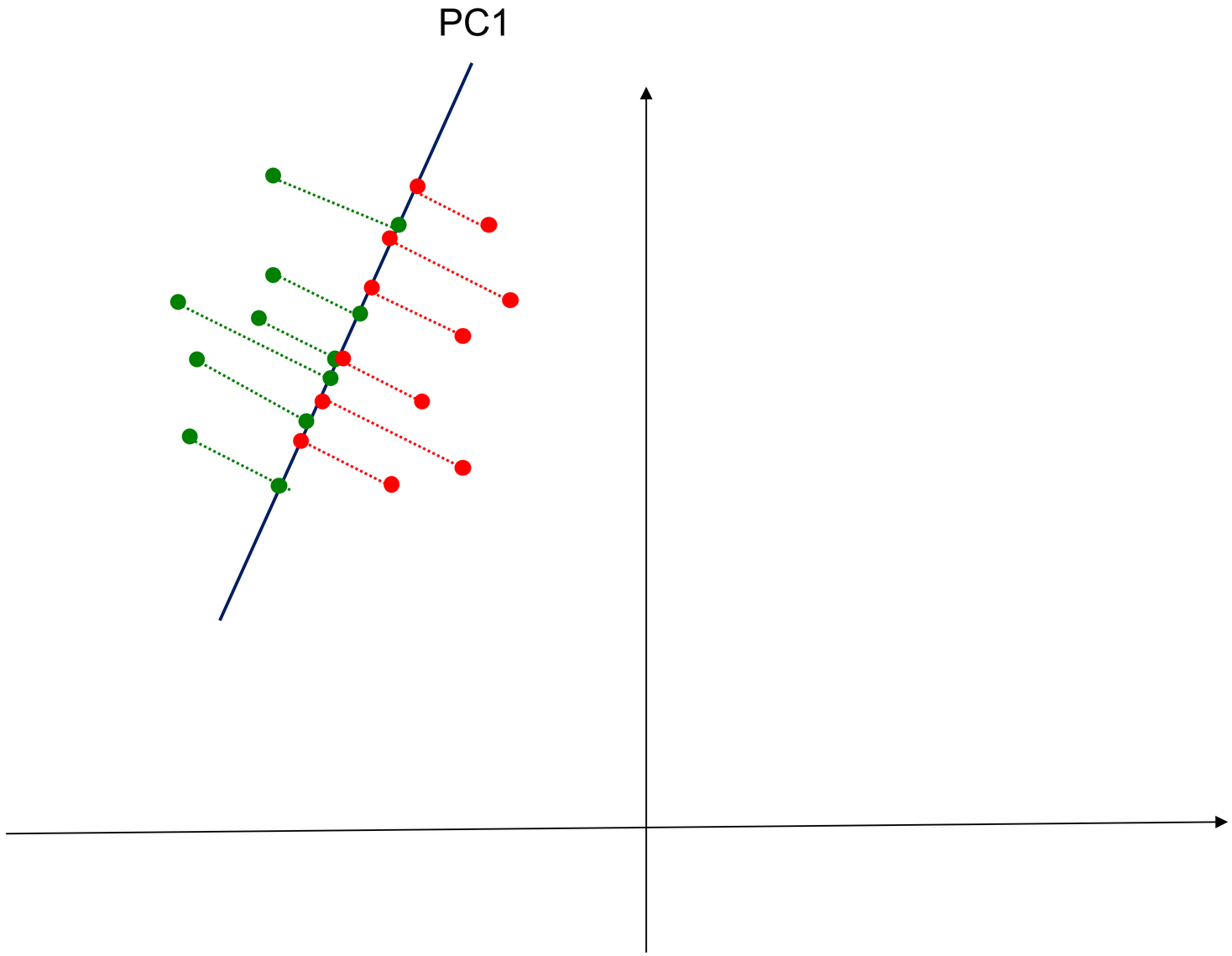
The GOAL of DA is the rotation of the Cartesian reference system along the directions of maximum variability AMONG GROUPS

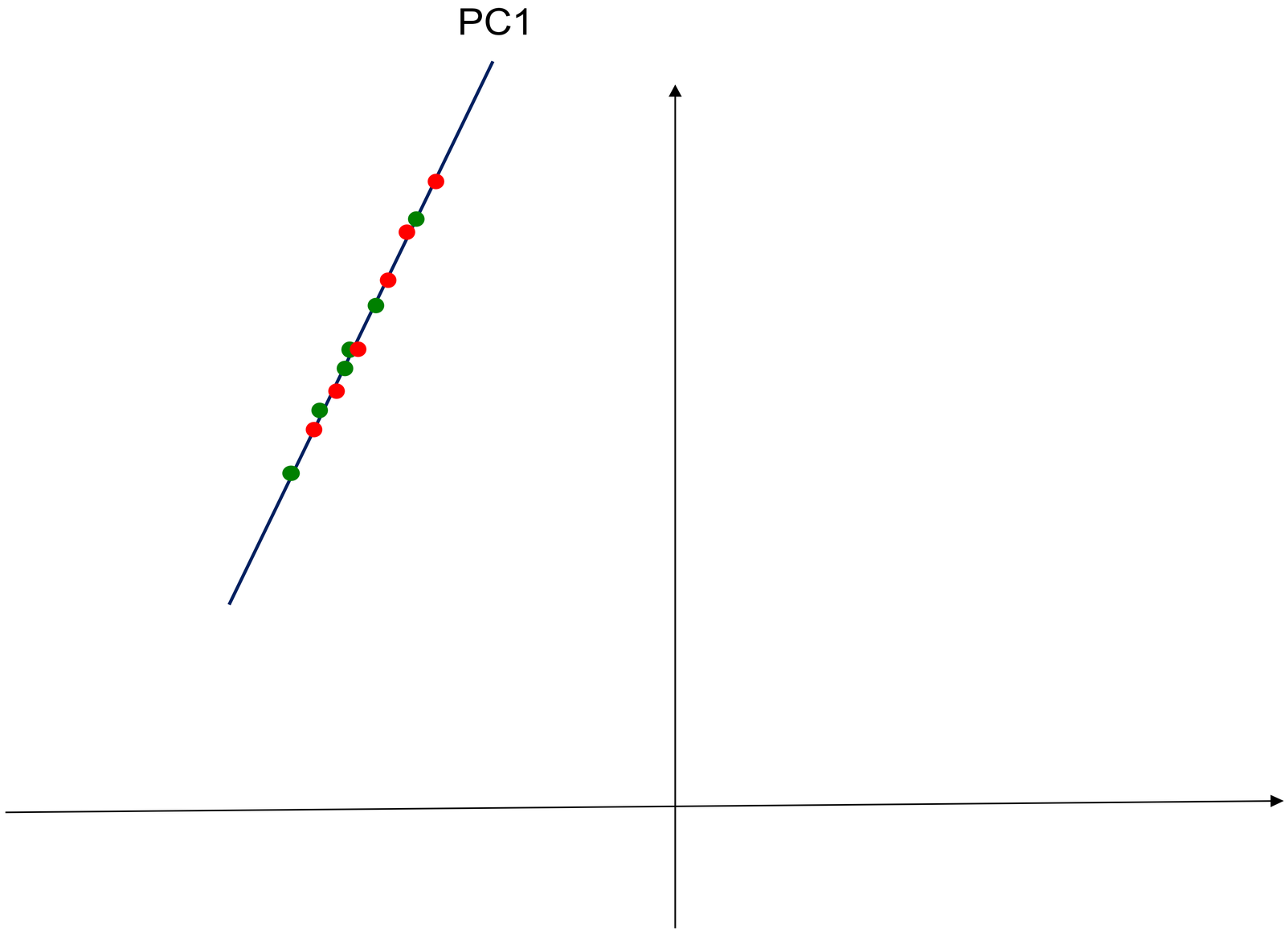
In PCA the starting point is the matrix SIGMA of DATA

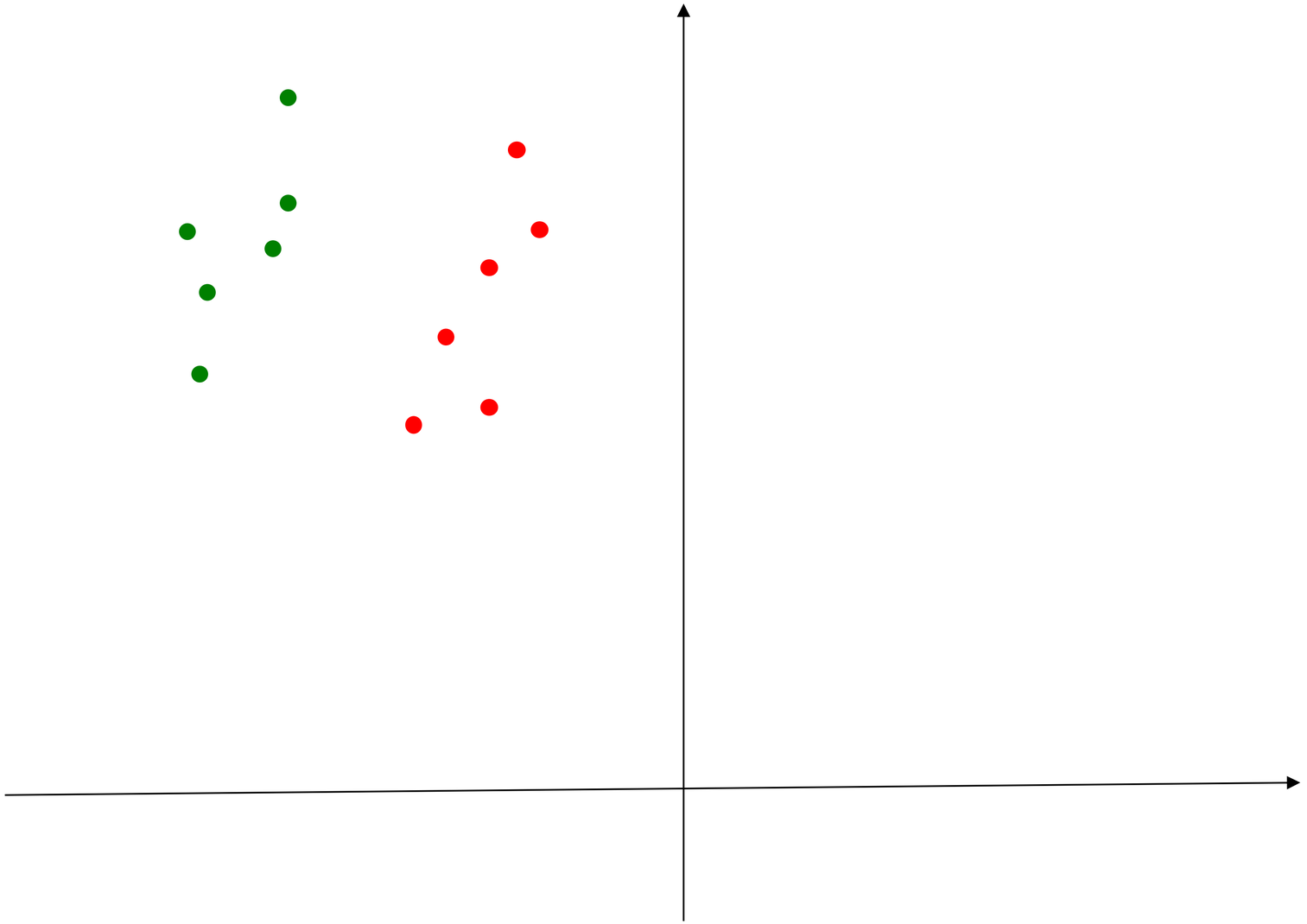
In DA the starting point is the matrix SIGMA of GROUPS

Suppose we have two groups of objects

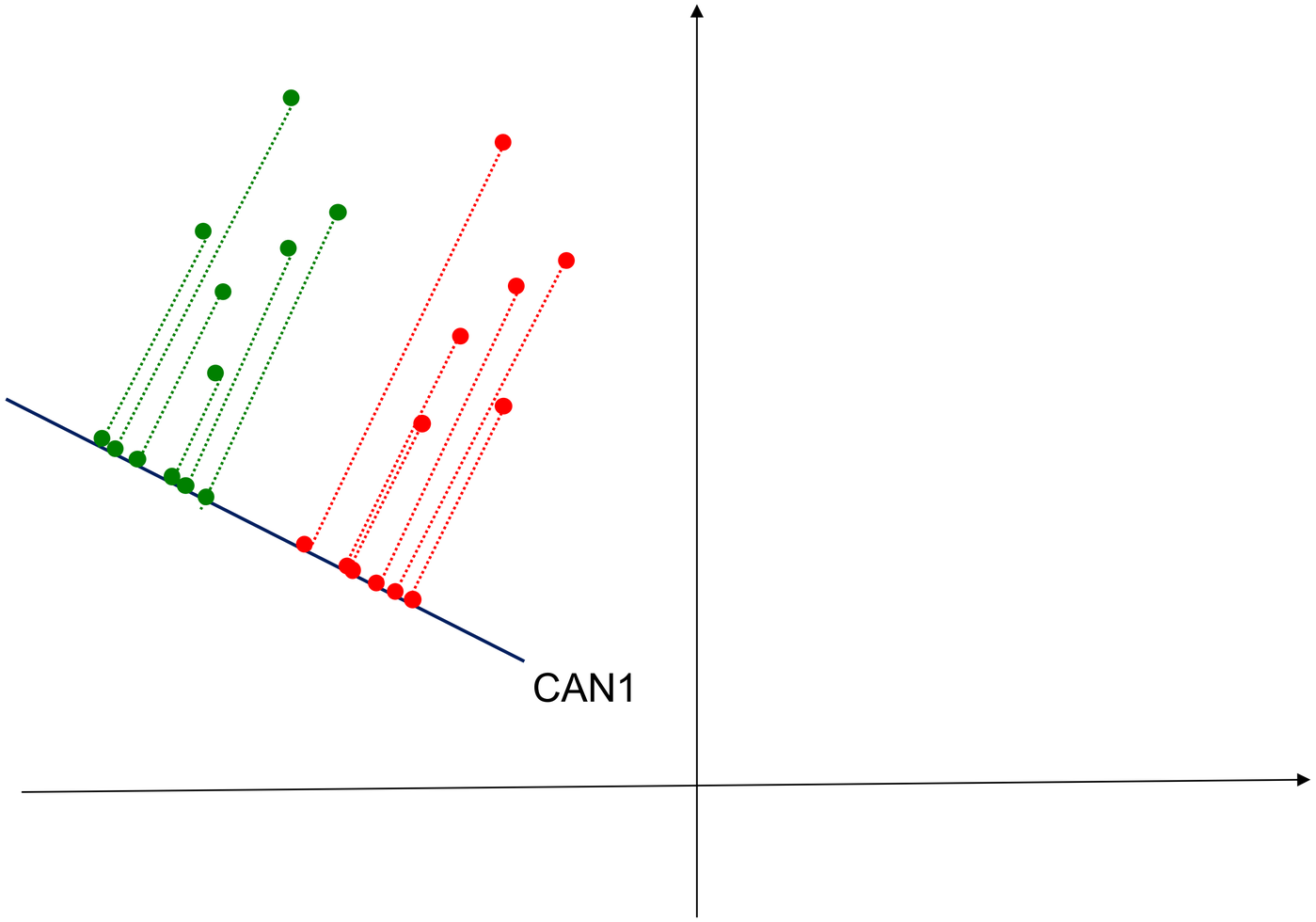



PC1

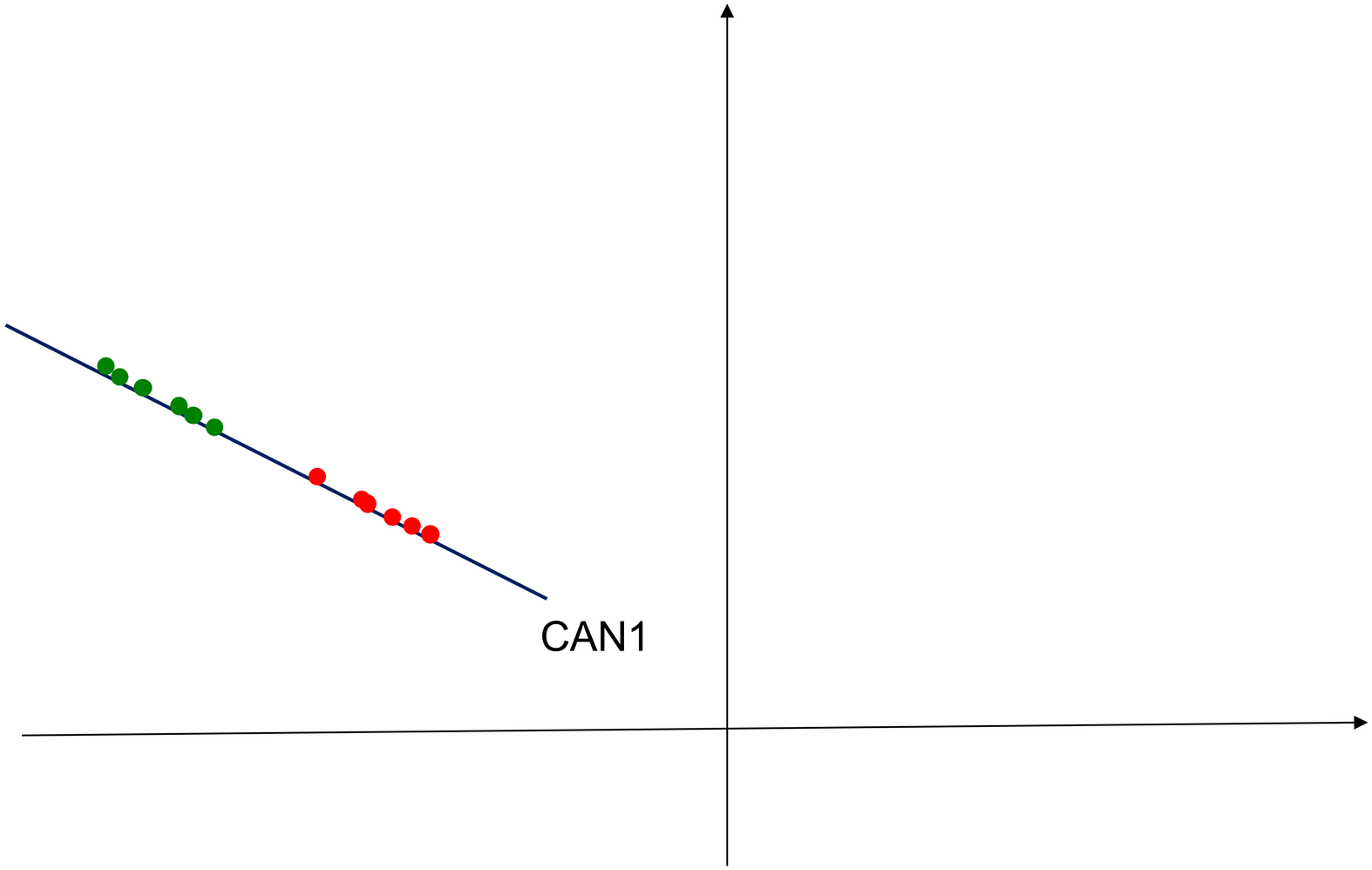

PC1

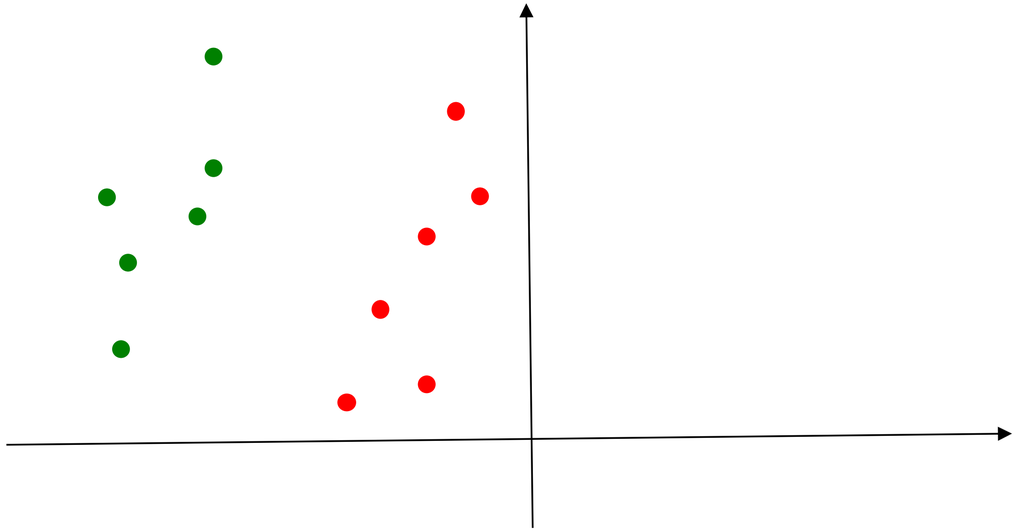


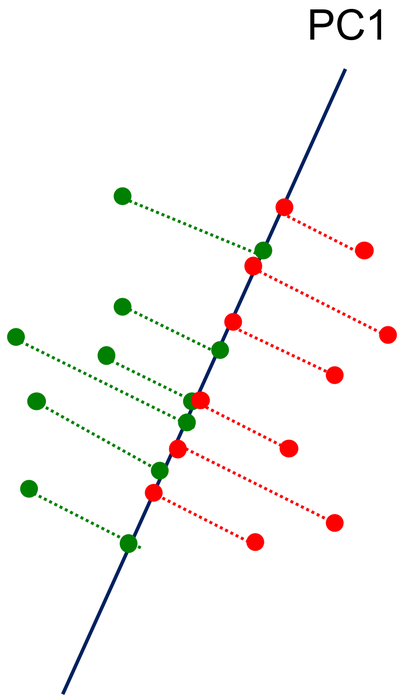
Discriminant Analysis identifies the right direction (CAN) on which to project the observations to highlight the groups

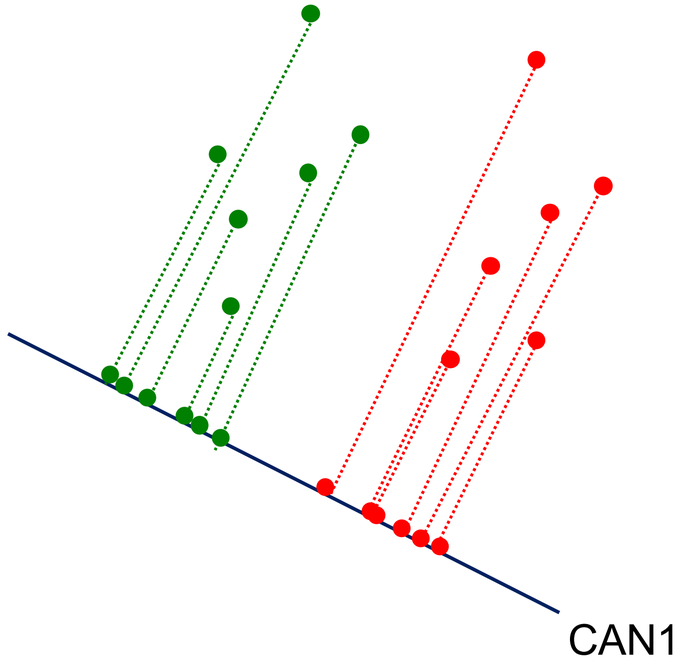


The rationale of DA is to seek, in the space of variables, the best directions able to highlight a clear separation between groups






PC1


CAN1

Main differences between PCA and DA

In PCA, we extract as many directions as the variables involved

In DA, k-1 directions are extracted, where k is the number of groups

PCA does not assign new observations

For DA, the assignment of new observations to groups is one of the main features

HOW DOES DA WORK?

Corrado Dimauro, Alberto Cesarani
Università di Sassari

Suppose we have p-variables

$$X_1, X_2, \ldots\ldots\ldots X_p$$

And k-groups

$$1,2,\ldots\ldots, \text{k-groups}$$

$$\text{With } k < p$$

Corrado Dimauro, Alberto Cesarani
Università di Sassari

The DA replaces the multiple variable of data X_1, X_2, ……..X_p with a single variable, called canonical function or discriminant function

$$\mathbf{CAN} = v_1X_1 + v_2X_2 + \ldots\ldots\ldots\ldots + v_pX_p$$

v_1, v_2, ……..,v_p **Canonical Coefficients**

v are the weights assigned to the original variables in the linear combination.

From X we will extract a particular variance covariance matrix

$$X = \begin{pmatrix} x_{11} & . & . & x_{1p} \\ . & . & . & . \\ . & . & . & . \\ x_{n1} & . & . & x_{np} \end{pmatrix}$$

Dim = ***n* x *p***

$$\Sigma = W^{-1} B$$

Dim = ***k*-1 × *k*-1**

The sum of squared deviations is the basic index to evaluate the variation in a particular dataset.

$$\sum_{i=1}^{n}(x_i - \bar{x})^2$$

The rationale of the DA is based on the ratio between: **B / W**

The sum of squared deviations BETWEEN groups (B)

The sum of squared deviations WITHIN groups (W)

The canonical coefficients v must be calculated under the condition that:

$$\mathbf{CAN} = v_1X_1 + v_2X_2 + \ldots\ldots\ldots\ldots + v_pX_p$$

the B matrix be as large as possible

And, at the same time,

the W matrix be as small as possible

In matrix algebra we have:

$$\lambda = \frac{v' B\, v}{v' W\, v}$$

Using the differential calculus, the ratio is maximum if the following equation is satisfied:

$$(\mathbf{W}^{-1}\mathbf{B} - \lambda\mathbf{I})\mathbf{v} = \mathbf{0}$$

This is a typical eigenvalue eigenvector problem

The number of eigenvectors we obtain is:

$$g = \min\left[(k-1), p\right]$$

To the largest eigenvalue corresponds to the first discriminant function and so on

So we have g Canonical functions (CANs)

The importance of the single CAN is evaluated with the following formula:

$$\frac{\lambda_h}{\sum \lambda_h}$$

How do we use the DA?

1) We must have a multivariate set of data, and objects must be divided in groups

2) The dataset is then divided in: training and validation datasets

3) DA is developed by using the training dataset and the extracted CANs are used to assign objects in the validation dataset

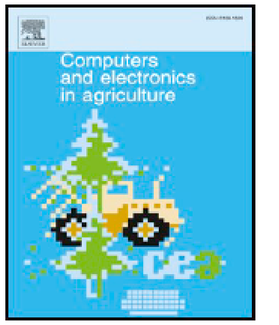
4) Performances of DA can be tested by comparing how DA assigns objects to groups with the real assignments

Computers and Electronics in Agriculture 176 (2020) 105657

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag



Original papers

Use of discriminant statistical procedures for an early detection of persistent lactations in dairy cows

Elisabetta Manca, Alberto Cesarani, Nicolò P.P. Macciotta, Alberto S. Atzori, Giuseppe Pulina, Corrado Dimauro*

Department of Agraria, University of Sassari, Viale Italia 39, 07100 Sassari, Italy

We had 2295 lactations

Primiparous cows
first parity group, FPG

Multiparous cows
multiple parity group, MPG

training **Validation**

training **Validation**

Each dataset contained two classes of lactations: LC and HC

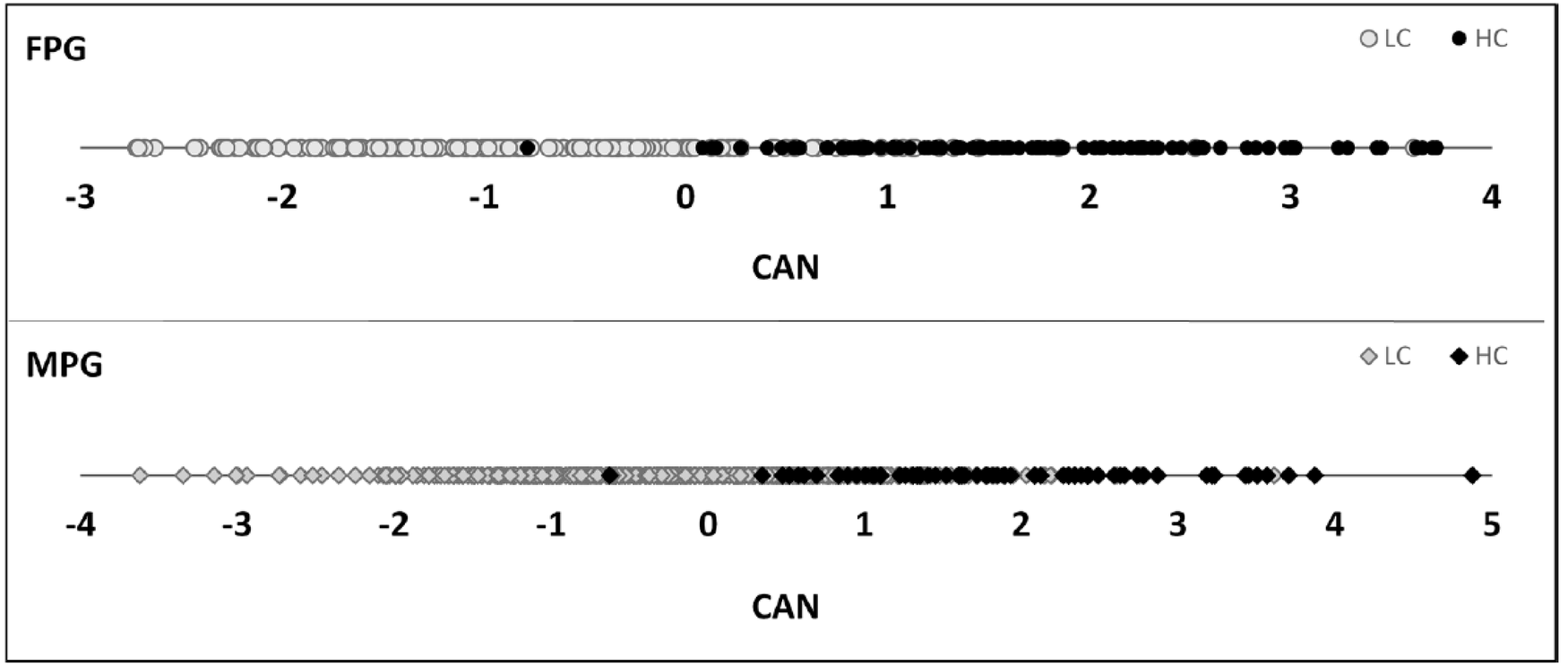


Fig. 2. Score plot of the canonical function (CAN), obtained by using the 4th degree polylomial, for first (FPG) and multiple (MPG) parity groups at 150 DIM between the low (LC) and high (HC) classes of production.

The discriminant analysis significantly separated LC from HC both for primiparous and multiparous cows.

The error percentage of lactations incorrectly assigned was 5% for primiparous and 7% for multiparous

Meat Science 87 (2011) 282–289

Contents lists available at ScienceDirect

Meat Science

journal homepage: www.elsevier.com/locate/meatsci

The volatile profile of *longissimus dorsi* muscle of heifers fed pasture, pasture silage or cereal concentrate: Implication for dietary discrimination

Valentina Vasta [a,*], Giuseppe Luciano [b], Corrado Dimauro [c], Florian Röhrle [d], Alessandro Priolo [a], Frank J. Monahan [d], Aidan P. Moloney [e]

[a] DACPA-Sezione di Scienze delle Produzioni Animali, University of Catania, via Valdisavoia 5, 95123 Catania, Italy
[b] Dottorato di Ricerca in Scienze delle Produzioni Animali, University of Catania, via Valdisavoia 5, 95123 Catania, Italy
[c] Dipartimento di Scienze Zootecniche, University of Sassari, Via De Nicola 9, 07100, Sassari, Italy
[d] School of Agriculture, Food Science and Veterinary Medicine, University College Dublin, Dublin 4, Ireland
[e] Teagasc, Animal and Grassland Research and Innovation Centre, Grange, Dunsany, County Meath, Ireland

Aim: study the effect of 4 different diets on the volatile profile of beef

Experimental design. **(32 heifrs)**

4 different diets (k=4): D1, D2, D3, D4.

8 heifers for each group (n=8);

95 volatile compounds

Methods.

Stepwise discriminant analysis

Discriminant analysis

Table 2
Correlations between the total canonical structure and the original variables.

Selected variables	Can1	Can2	Can3
1-Butanol, 2-methyl	−0.44	0.33	−0.12
2-Buten-1-ol	0.37	−0.35	−0.30
3-Undecanone	0.66	0.23	−0.17
Pinane	0.15	−0.05	0.01
Unidentified terpene (LRI 1023)	−0.20	0.53	−0.13
Pentanale	−0.13	−0.09	−0.52
Decanal	−0.03	−0.11	0.03
Skatole	0.47	−0.37	0.01
Cuminic alcohol	0.36	0.11	−0.16
Germacrene D	0.20	−0.29	0.15
1-Hexene	0.38	0.10	0.09
Unidentified compound (LRI 1013)[a]	0.24	0.17	−0.00
Dodecanal	0.12	0.17	0.08
Benzene ethenyl	0.06	−0.35	0.10
Acetaldehyde phenyl	−0.00	−0.11	0.13
Unidentified compound (LRI 1083)[b]	0.28	−0.32	0.19
Eigenvalue	78.13	19.18	3.06
Variance explained (%)	77.8	19.2	3.0
Cumulative variance (%)	77.8	97.0	100.0

[a] Spectrum list 55 (100), 70 (93), 83 (67), 97 (16), 108 (5), 111 (4).
[b] Spectrum list: 93 (100), 193 (79),126 (77), 117 (71), 179 (61), 135 (28).

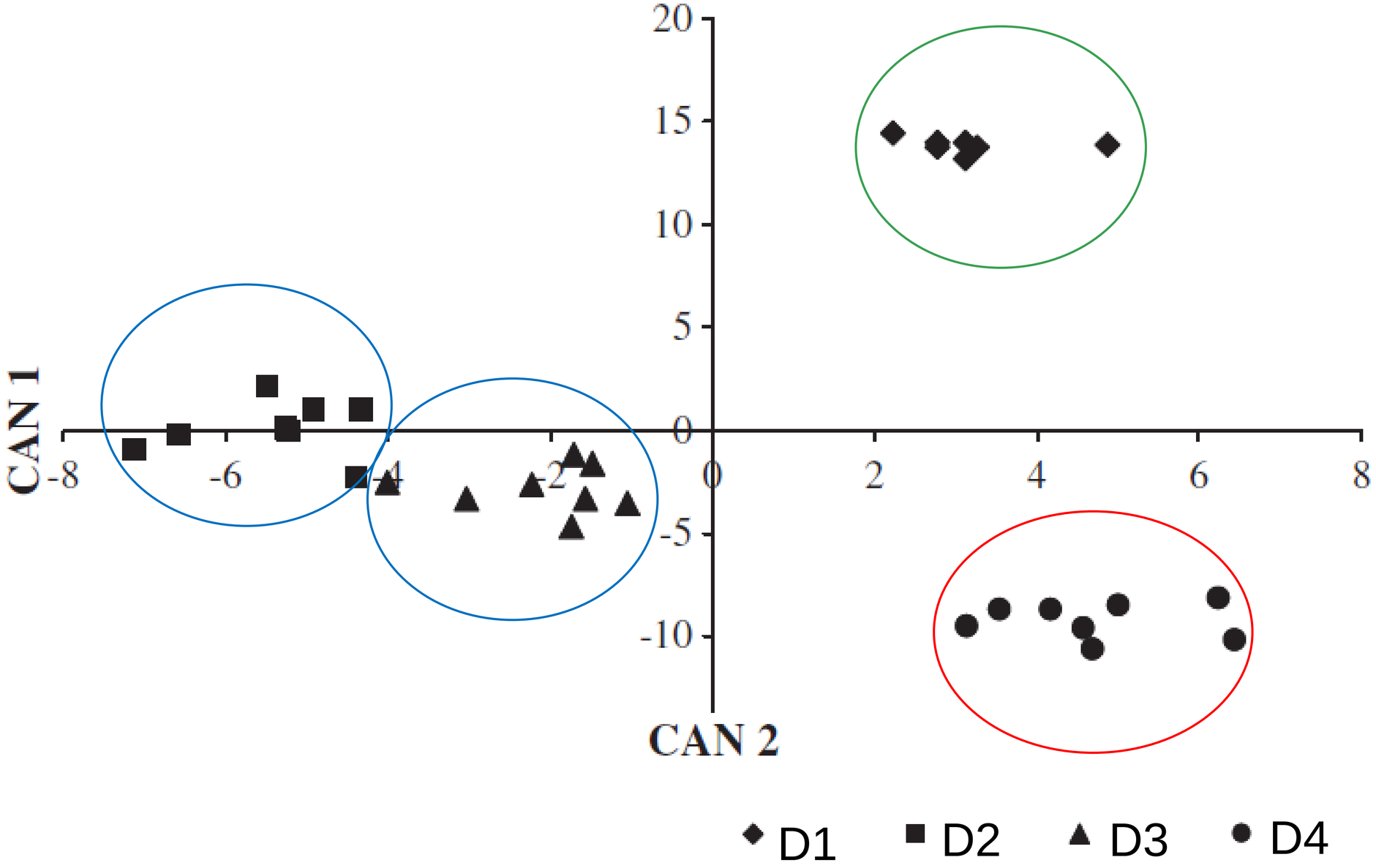


Fig. 2. Discrimination of dietary treatment on the basis of canonical discriminant analysis.

All animals were correctly assigned to groups

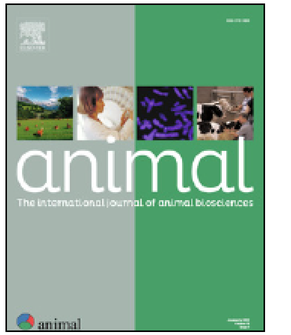
Table 3
Mahlanobis distances and significances (in bold) between the four diets.

	D1	D2	D3	D4
D1		**<0.0001**	**<0.0001**	**<0.0001**
D2	267		**0.008**	**<0.0001**
D3	315	38		**<0.0001**
D4	541	191	99	

Animal 18 (2024) 101102

Contents lists available at ScienceDirect

Animal
The international journal of animal biosciences



Predicting feed efficiency of Angus steers using the gastrointestinal microbiome

M. Congiu [a,b], J. Lourenco [b], A. Cesarani [a,b,*], U. Lamichhane [b], N.P.P. Macciotta [a], C. Dimauro [a]

[a] *Dipartimento di Agraria, University of Sassari, Sassari 07100, Italy*
[b] *Department of Animal and Dairy Science, University of Georgia, Athens 30602, GA, USA*

In this paper, a statistical procedure to classify Angus steers in two groups with extreme Residual Feed Intake (RFI) values, using their microbiota profile, was developed. In particular the fecal microbiome was used

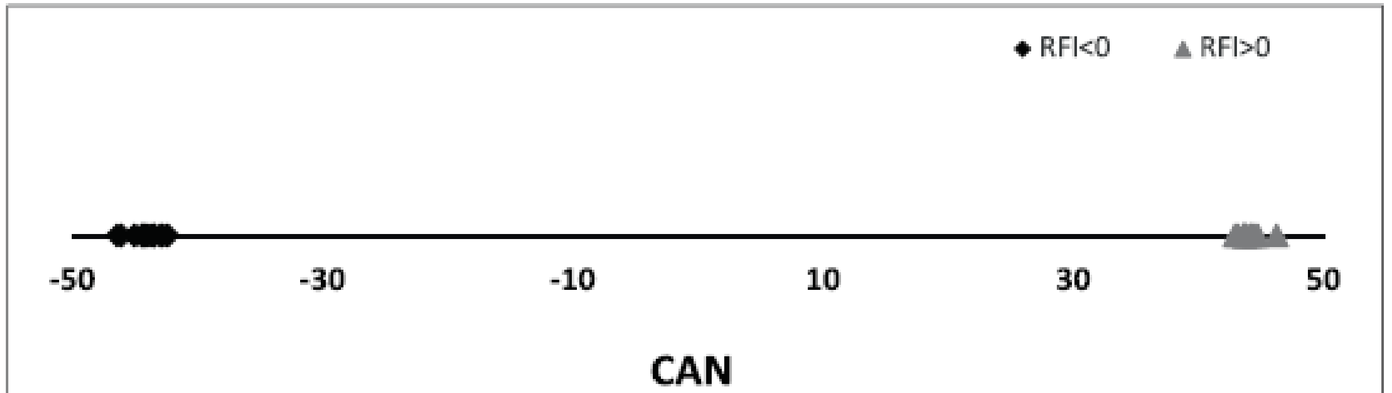
Fecal samples were collected at weaning. A total of 119 bacterial families (BFs) were retrieved from the fecal samples.

The DA was used to investigate whether BFs were able to distinguish between animals with positive or negative RFI values.

The DA using 18 BFs selected from the stepwise was able to correctly assign all animals to the proper RFI groups.

Table 4
List of the 18 bacterial families, with their absolute standardized canonical coefficient (CC) and relative abundances, able to discriminate between Angus steers with negative (NRFI) and positive (PRFI) residual feed intake.

Bacterial Families	Standardized CC	Relative abundances	
		NRFI	PRFI
Rhizobiaceae	−403	0.01	0.00
Erysipelatoclostridiaceae	−244	0.05	0.09
Acidaminococcaceae	−170	1.04	0.31
p_2534_18B5_gut_group	−161	0.70	0.09
Fibrobacteraceae	−108	0.05	0.00
Beijerinckiaceae	−107	0.02	0.01
Lachnospiraceae	−61	6.64	10.78
Enterobacteriaceae	−49	0.2	0.67
Prevotellaceae	−34	8.92	5.53
WCHB1_41	−31	0.09	0.05
Moraxellaceae	−25	0.01	0.02
Campylobacteraceae	−25	0.28	0.07
Oscillospiraceae	31	17.17	20.05
Acholeplasmataceae	41	0.03	0.05
RF39	61	0.02	0.09
Succinivibrionaceae	206	0.00	0.01
Atopobiaceae	294	0.05	0.19
Comamonadaceae	340	0.00	0.02



ANIMAL GENETICS Immunogenetics, Molecular Genetics and Functional Genomics

doi: 10.1111/age.12021

Use of the canonical discriminant analysis to select SNP markers for bovine breed assignment and traceability purposes

C. Dimauro*, M. Cellesi*, R. Steri*, G. Gaspa*, S. Sorbolini*, A. Stella† and N. P. P. Macciotta*

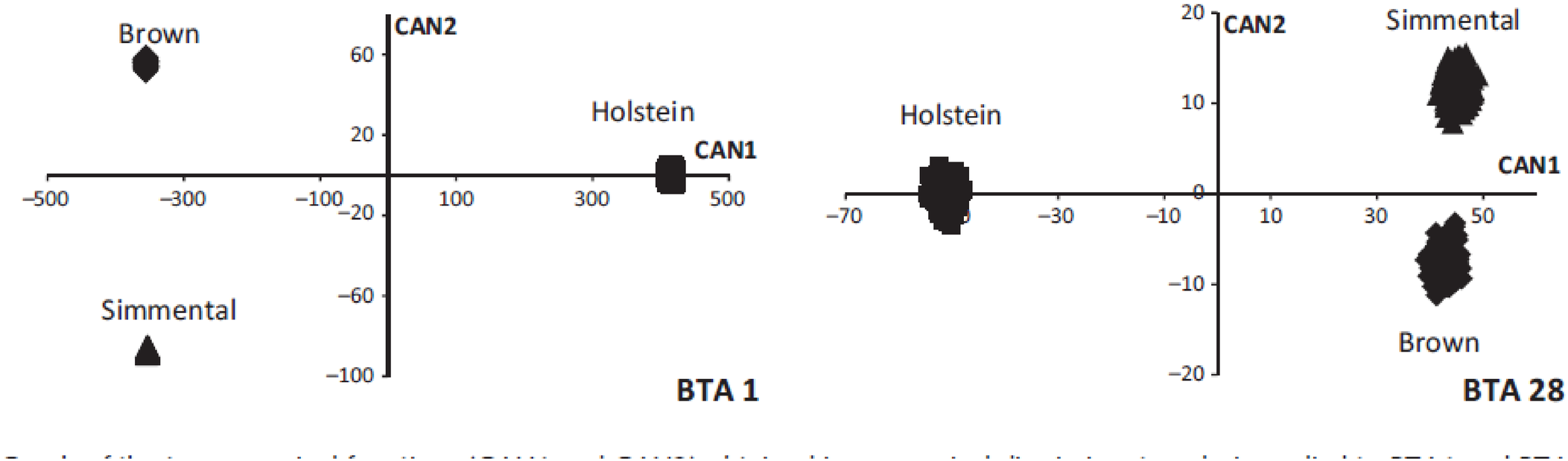
*Dipartimento di Agraria, Università di Sassari, Via De Nicola 9, 07100, Sassari, Italy. †Istituto di biologia e biotecnologia agraria CNR, -20133, Milano, Italy.

The data came from 1042 Holstein, 750 Brown Swiss, and 480 Simmental bulls genotyped using the Illumina 50K BeadChip

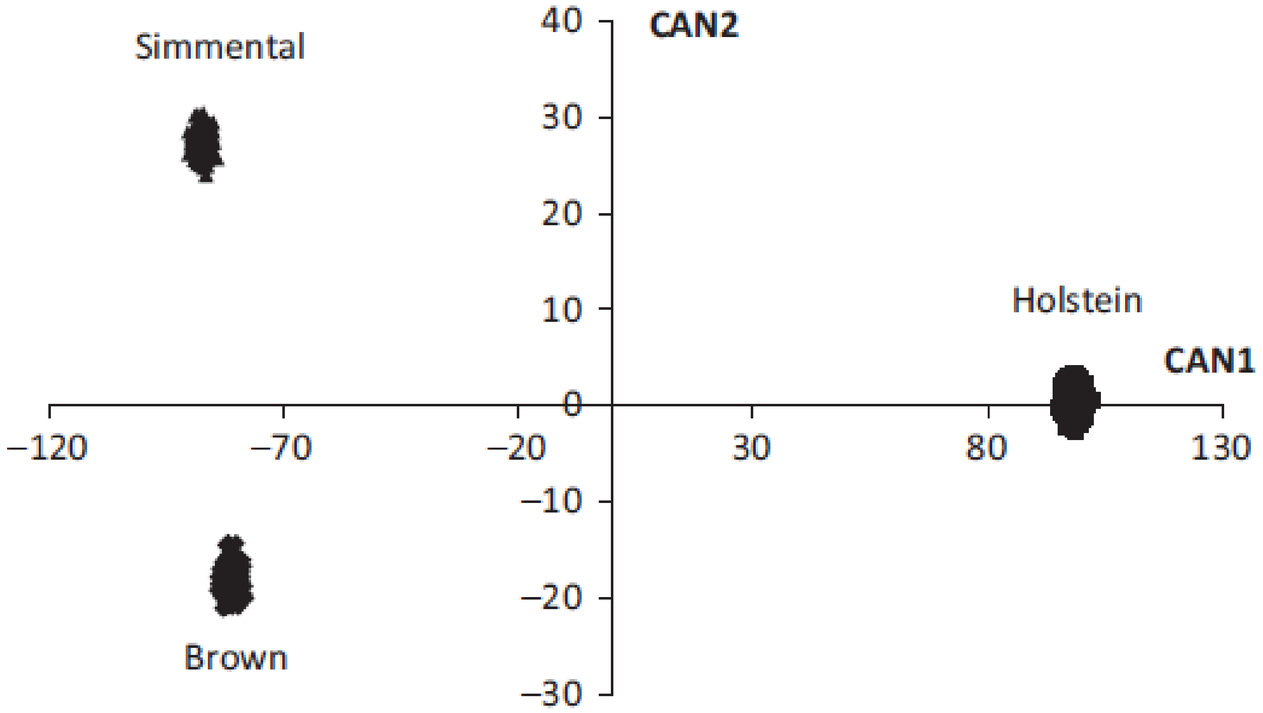
The discriminant analysis was used to discriminate the three breeds and assign new observations to the correct group.

The matrix of data consisted of more than 30K SNP variables and 2K animals. In this condition, multivariate techniques became meaningless because the number of rows (the animals) is lower than the number of variables (the SNP)

So, to at least partially overcome this problem, statistical analyses were developed by chromosome and, in each chromosome, the SDA was used to further reduce the number of SNP-variables



SNPs selected in each chromosome were joined and a new run of the SDA was developed to select the final maximum number of SNPs able to discriminate breeds. This number is lower than the number of involved animals (< 2272)



GW-DA with 1836 SNPs

In a new run of the SDA, the minimum number of markers able to highlight the existence of the groups was obtained. At the end, 48 SNPs were retained and used in a new GW-CDA

	Brown	Simmental
Simmental	301 (<0.0001)	
Holstein	4300 (<0.0001)	3574 (<0.0001)

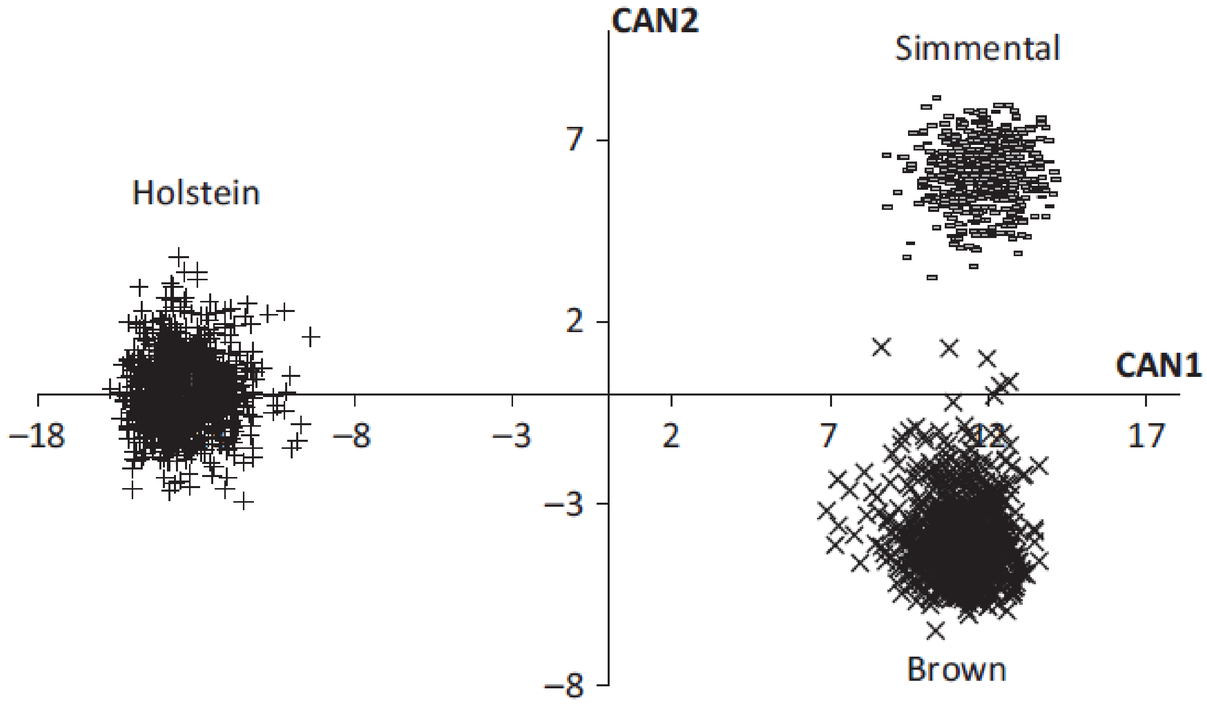


Figure 3 Graph of the two canonical functions (CAN1 and CAN2) obtained in a genome-wide canonical discriminant analysis using a restricted number (48) of linearly independent SNP variables.

Table 2 Canonical coefficients (CC), in the two canonical functions (CAN1 and CAN2), for the most 48 discriminant markers selected among SNPs belonging to the Illumina BovineSNP50 v2 BeadChip.

SNP name	BTA	CC (CAN1)	SNP name	BTA	CC (CAN2)
BTB-01524285	5	0.944	*Hapmap56688-rs29025335*	6	−0.671
ARS-BFGL-NGS-116089	15	0.941	*ARS-BFGL-NGS-100916*	6	−0.666
Hapmap51971-BTA-18711	11	0.936	*ARS-BFGL-NGS-103634*	18	−0.664
BTB-01648149	3	0.936	*Hapmap30962-BTC-032558*	6	−0.651
BTA-23857-no-rs	12	0.933	*ARS-BFGL-NGS-41271*	20	−0.648
BTB-01267305	5	0.932	*ARS-BFGL-NGS-108820*	6	−0.645
BTA-73563-no-rs	5	0.931	*BTB-00049653*	1	−0.640
BTA-79188-no-rs	1	0.930	*Hapmap27224-BTA-161106*	6	−0.640
ARS-BFGL-NGS-3048	29	0.929	*ARS-BFGL-NGS-67658*	6	−0.640
BTB-00498059	12	0.928	*BTB-00259302*	6	−0.639
Hapmap33485-BTA-144281	6	0.928	*Hapmap54879-rs29017018*	6	−0.635
Hapmap55512-rs29011234	26	0.928	*Hapmap52160-rs29020798*	6	−0.627
ARS-BFGL-NGS-22403	16	−0.921	*ARS-BFGL-NGS-20141*	7	0.633
BTA-58999-no-rs	24	−0.922	*BTA-37834-no-rs*	5	0.636
UA-IFASA-3757	13	−0.922	*BTA-110240-no-rs*	6	0.636
BTB-00506196	12	−0.922	*Hapmap42715-BTA-87995*	6	0.643
BTB-00951350	27	−0.925	*Hapmap57799-rs29012894*	11	0.643
BTB-00506214	12	−0.926	*ARS-BFGL-BAC-33135*	18	0.650
ARS-BFGL-NGS-36907	26	−0.928	*Hapmap50117-BTA-81807*	6	0.650
BTB-00146014	3	−0.928	*Hapmap44452-BTA-22099*	6	0.681
Hapmap44270-BTA-67318	9	−0.928	*Hapmap33128-BTC-041916*	6	0.766
BTB-00178642	4	−0.928	*Hapmap26269-BTC-041695*	6	0.782
BTA-18115-no-rs	2	−0.937	*ARS-BFGL-NGS-38827*	6	0.785
Hapmap51008-BTA-62521	27	−0.943	*Hapmap27692-BTC-042876*	6	0.787