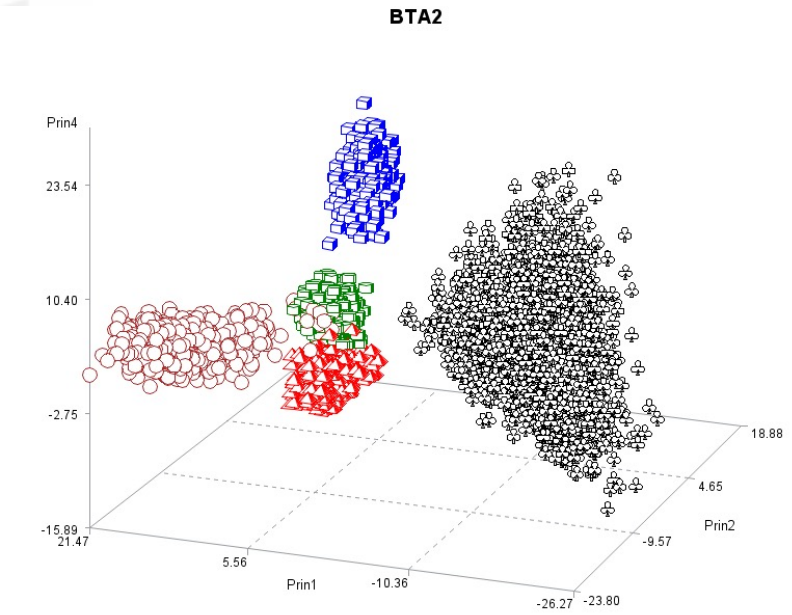
# Principal component analysis (PCA)



# Main aims of multivariate statistics

- ✓ Reduction of system dimensions
- ✓ Search for latent structures
- ✓ Analysis of dependence relatioships

# *Reduction of system dimensions*

- ✓ **Reduction from p to q variables (q<p)**
- ✓ **Limiting loss of information**
- ✓ **New variables with technical or biological meaning**

# *Principal Component Analysis*

- ✓ **Dimension reduction method**
- ✓ **Developed by K. Pearson and H. Hotelling**
- ✓ **Used in social, economic, and life sciences**
- ✓ **Model-free approach**

# PCA under an analytic perspective

$$X = \begin{bmatrix} X_{11} & X_{12} & \cdots\cdots & X_{1p} \\ X_{21} & X_{22} & \cdots\cdots & X_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ X_{N1} & X_{N2} & \cdots\cdots & X_{Np} \end{bmatrix}_{(Nxp)}$$

**Data matrix**

# PCA under an analytic perspective

$$\mathbf{Z} = \begin{bmatrix} z_{11} & z_{12} & \cdots\cdots & z_{1p} \\ z_{21} & z_{22} & \cdots\cdots & z_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ z_{N1} & z_{N2} & \cdots\cdots & z_{Np} \end{bmatrix}_{(Nxp)}$$

**Standardized data matrix**

# PCA under an analytic perspective

$$\mathbf{R} = \begin{bmatrix} r_{11} & r_{12} & \cdots\cdots & r_{1p} \\ r_{21} & r_{22} & \cdots\cdots & r_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ r_{p1} & r_{p2} & \cdots\cdots & r_{Np} \end{bmatrix}_{(Nxp)}$$

**Correlation matrix**

# PCA under an analytic perspective

## Vector tansformation

Given a column vector $\mathbf{x}_{n,1}$ and a matrix $\mathbf{A}_{m,n}$
The product

$$\mathbf{y=Ax}$$

defines a transformation, creating a vector **y** with the same number of rows of the matrix **A**

i.e. the sistem coordinates are transformed from the n-dimensions of the space of **x** to the m-dimensions of the space of **A**

# PCA under an analytic perspective

The first principal component ($Y_1$) of a multivariate system of $X_1$, $X_2$,…..Xp variables is the linear combination

$$\mathbf{Y_1}=a_{11}\boldsymbol{X_1} + a_{21}\boldsymbol{X_2}+\ldots+a_{p1}\boldsymbol{X_p}$$

$a_{i1}$ = coefficients of the eigenvector $\mathbf{a}_1$ associated to the largest eigenvalue $\lambda_1$ of **R**

constraints

$\mathbf{a}_1'\mathbf{a}_1 = \mathbf{1}$ normalised vector

var($\boldsymbol{Y_1}$) = max. Maximum variance

# PCA under an analytic perspective

The second principal component ($Y_2$) of a multivariate system of $X_1$, $X_2$,…..Xp variables is the linear combination

$$\mathbf{Y_2}=a_{12}\mathbf{X_1} + a_{22}\mathbf{X_2}+\ldots+a_{p2}\mathbf{X_p}$$

$a_{i2}$ = coefficients of the eigenvector $\mathbf{a}_2$ associated to the second largest eigenvalue $\lambda_2$ of **R**

constraints

| | |
|---|---|
| $\mathbf{a}_2'\mathbf{a}_2 = \mathbf{1}$ | normalised vector |
| $\mathbf{a}_1'\mathbf{a}_2 = \mathbf{0}$ | orthogonality |
| var($\boldsymbol{Y_2}$) = max. | Maximum variance |

# PCA under an analytic perspective

The j-th principal component ($Y_j$) of a multivariate system of $X_1$, $X_2$,…..Xp variables is the linear combination

$$\mathbf{Y_j}=a_{1j}\boldsymbol{X_1} + a_{2j}\boldsymbol{X_2}+\ldots+a_{pj}\boldsymbol{X_p}$$

$a_{ij}$ = coefficients of the eigenvector $\mathbf{a}_j$ associated to the j-th largest eigenvalue $\lambda_j$ of **R**

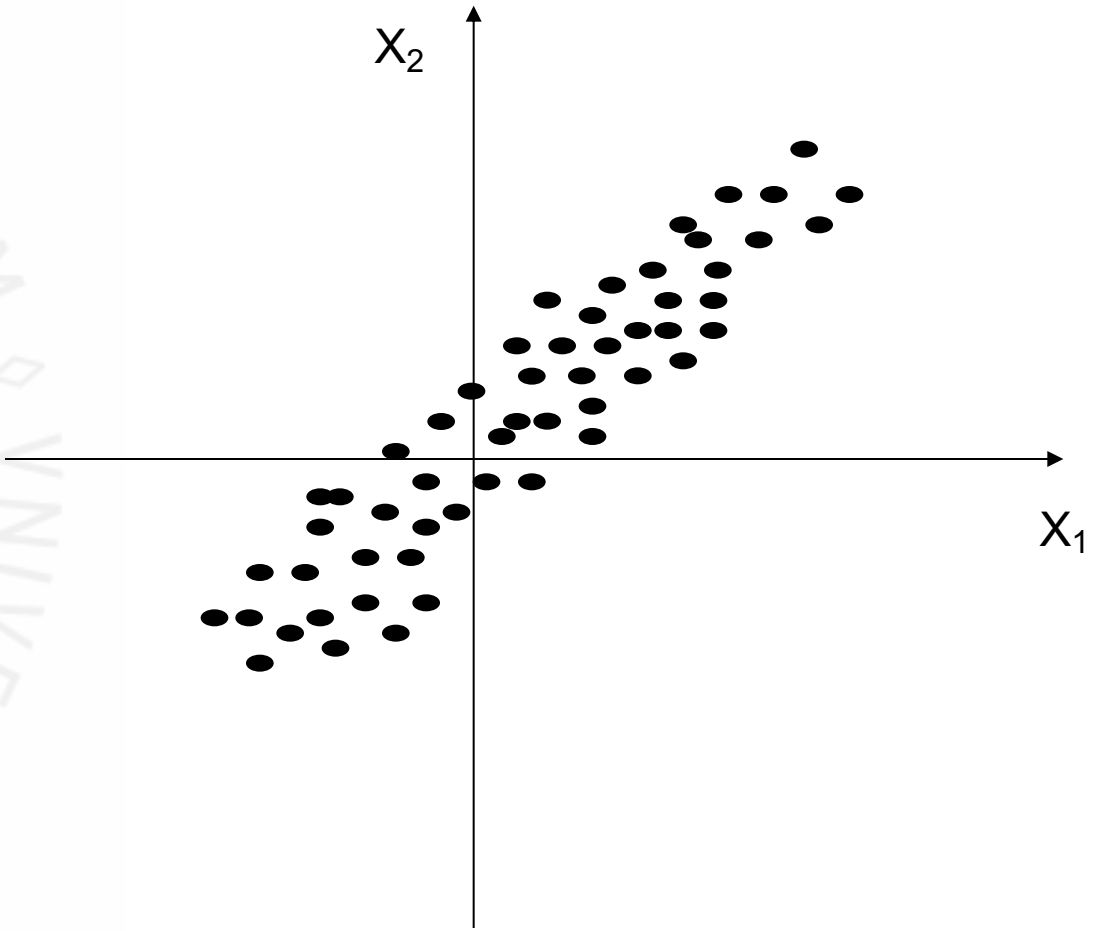
$\lambda_j$ is the variance of the j-th component

$$\frac{\lambda_j}{tr(\mathbf{R})} = \frac{\lambda_j}{p}$$ amount of variance explained by the j-th component

# PCA under an analytic perspective

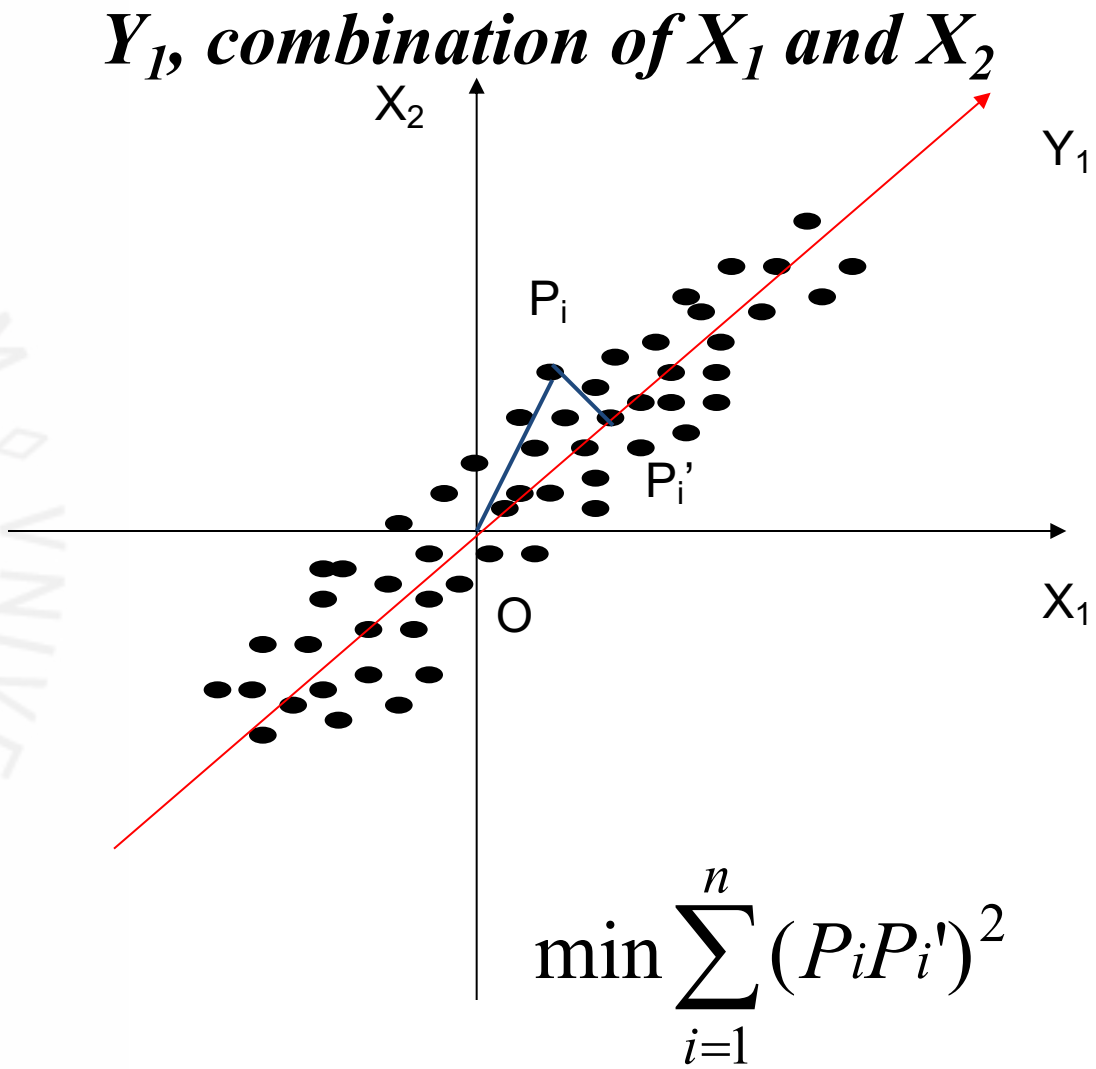
- ✓ Linear combinations of p original variables
- ✓ PC coefficients are the elements of the eigenvectors associated to the eigenvalues of the correlation matrix **R** of original variables
- ✓ PC are extracted in order to orthognally partition the variance in sequentially smaller portions
- ✓ PC are orthogonal
- ✓ The number of extracted PC is equal to the number of the original variables

# PCA under a geometric perspective




•System defined by two variables (X1 e X2)

# PCA under a geometric perspective




***Optimal rotation***

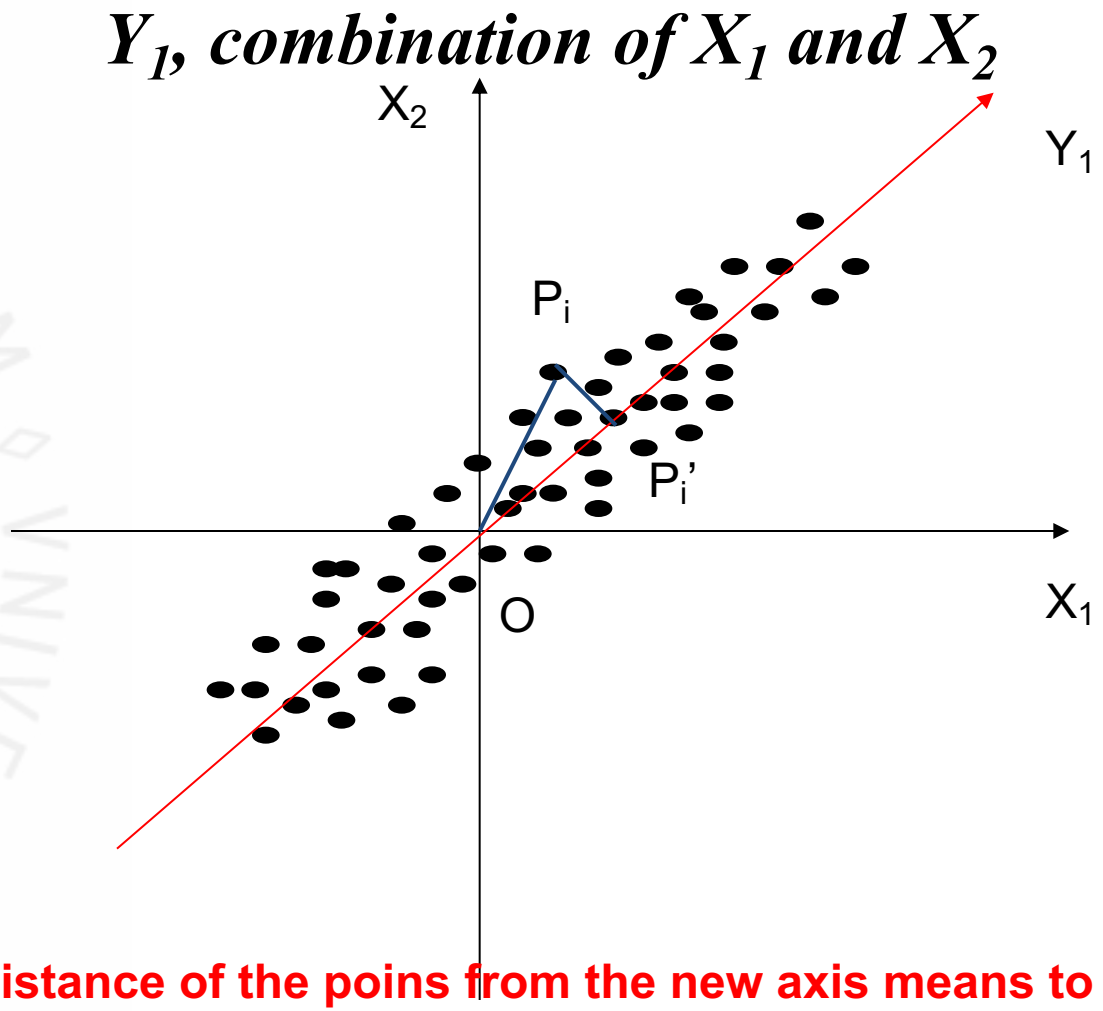
# PCA under a geometric perspective

***Applying the Pitagora's theorem to the whole set o f points***

$$\frac{1}{n-1}\sum_{i=1}^{n}(OP_i)^2 = \frac{1}{n-1}\sum_{i=1}^{n}(OP_i')^2 + \frac{1}{n-1}\sum_{i=1}^{n}(P_iP_i')^2$$

***The first term on the RHS, i.e. the sampling variance, has to be maximized***

# PCA under a geometric perspective




**Minimising the distance of the poins from the new axis means to find the linear combination *OY1* that allows to the projected points the maximum variance**

# PCA under a geometric perspective

- ✓ Creation of new variables that are linear combination of original variables
- ✓ Rotation of the reference axes
- ✓ The new axes are oriented in the sense of maximum explained variance

# PCA and matrix algebra

A symmetric matrix **S** can be written in the form:

$$\mathbf{S} = \mathbf{PD}(\lambda i)\mathbf{P}'$$

- If **P** is equal to the eigenvector matrix (orthogonal)

$$\mathbf{S} = \mathbf{PD}\left(\sqrt{\lambda_j}\right)\mathbf{D}\left(\sqrt{\lambda_j}\right)\mathbf{P}'$$

# PCA and matrix algebra

If we put $$L = \mathbf{PD}\left(\sqrt{\lambda_j}\right)$$

**L** columns reproduce **S** with the formula:

$$\mathbf{S=LL'=}\ \lambda_1\mathbf{a}_1\mathbf{a}_1' + \ldots\ldots\ldots\ \lambda_p\mathbf{a}_p\mathbf{a}_p'$$

PC analysis is equivalent to the factorization of **S** into a product of a matrix **L** for its transposed

# *PCA as factorisation of S*

- ✓ PC analysis is equivalent to the factorization of **S** into a product of a matrix **L** for its transposed
- ✓ In PCA the factorisation of **S** is **unique**
- ✓ PC coefficients are chosen in order to partition the total variance orthogonally in subsequent smaller amounts

# *Meaning and structure of PCA*

- Each $Y_j$ is called **principal component** (PC) of the system.

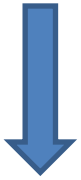
- Values of $a_{ij}$ are principal component **coefficients**

- Values of $y_i$ calculated for each observation are principal component ***scores***, i.e. the coordinates of the observations in the new axis.

# Meaning and structure of PCA

$$\mathbf{Z} = \begin{bmatrix} z_{11} & z_{12} & \cdots\cdots & z_{1p} \\ z_{21} & z_{22} & \cdots\cdots & z_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ z_{N1} & z_{N2} & \cdots\cdots & z_{Np} \end{bmatrix}_{(Nxp)}$$

**Za**



**From data matrix to score matrix**

$$\mathbf{Y} = \begin{bmatrix} Y_{11} & Y_{12} & \cdots\cdots & Y_{1p} \\ Y_{21} & Y_{22} & \cdots\cdots & Y_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ Y_{N1} & Y_{N2} & \cdots\cdots & Y_{Np} \end{bmatrix}_{(Nxp)}$$

# *Which PC shoud be retained?*

- ✓ The number of extracted PC is equal to the number of the original variables p
- ✓ A smaller number q (q<p) of PC> should be retained

**1) Amount of explained variance**

**2) Interpretation of the PC structure**

# Explained variance

$$\frac{\sum \lambda_j}{tr(\mathbf{R})} = \frac{\sum \lambda_j}{p}$$

- ✓ Usually the first 4 or 5 PC are able to explain a large amount of variance (>75%)
- ✓ The first PC often explains a large amount of variance but does not show a defined structure
- ✓ Subsequent PC may show more interesting features

# Explained variance

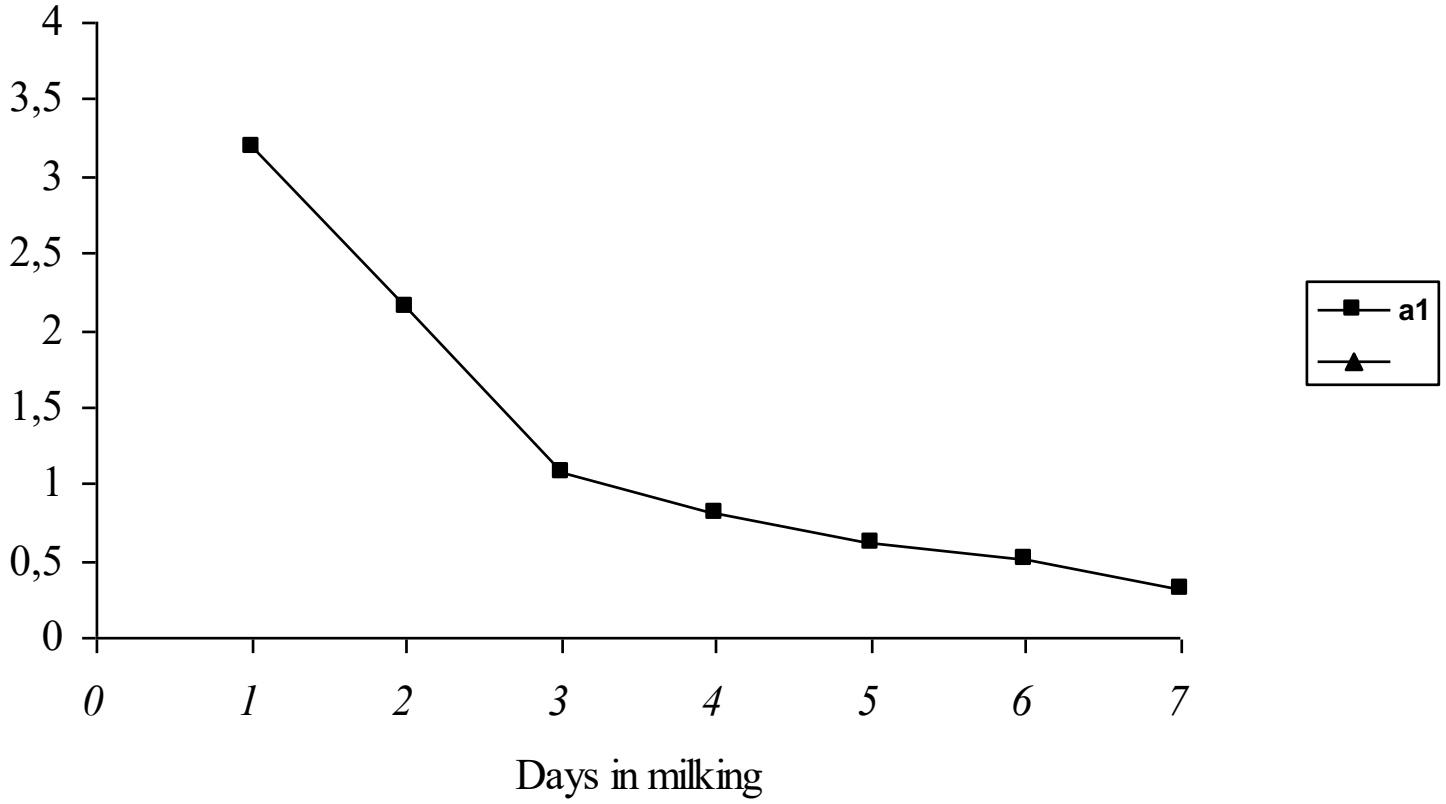
Eigenvalues of the correlation matrix

| | Eigenvalue | Proportion | Cumulative |
|---|---|---|---|
| 1 | 3.19091616 | 0.3545 | 0.3545 |
| 2 | 2.14875056 | 0.2388 | 0.5933 |
| 3 | 1.06869307 | 0.1187 | 0.7120 |
| 4 | 0.81066724 | 0.0901 | 0.8021 |
| 5 | 0.60729098 | 0.0675 | 0.8696 |
| 6 | 0.51249382 | 0.0569 | 0.9265 |
| 7 | 0.32280601 | 0.0359 | 0.9624 |
| 8 | 0.25248232 | 0.0281 | 0.9905 |
| 9 | 0.08589984 | 0.0095 | 1.0000 |

$$\lambda_j \qquad \frac{\lambda j}{p} \qquad \frac{\sum \lambda_j}{p}$$

***The first four PC explain about 80% of the variance***

# Explained variance




- ✓ The partition of variance is easier to understand by looking at the eigenvalue plot

# PC interpretation

Eigenvector matrix **A**

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots\cdots & a_{1p} \\ a_{21} & a_{22} & \cdots\cdots & a_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ a_{p1} & a_{p2} & \cdots\cdots & a_{pp} \end{bmatrix}_{(Nxp)}$$

# PC interpretation

- ✓ Sign and value of of eigenvector coefficients $a_{ij}$ represent the direction and the importance of the contribution of the *i-th* original variable on the *j-th* princpal component

- ✓ If principal components are extracted from the correlation matrix **R**, corelations between original variables and the *j-th* principal component are given by the vector

$$\mathrm{a}_j \sqrt{\lambda_j}$$

# PC interpretation

## Eigenvector matrix A

| Variable | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a_5$ | $a_6$ | $a_7$ | $a_8$ | $a_9$ |
|---|---|---|---|---|---|---|---|---|---|
| Milk | -0,20 | **0,48** | -0,09 | 0,26 | 0,34 | 0,61 | -0,35 | 0,20 | -0,03 |
| Fat | 0,01 | -0,22 | **0,86** | -0,15 | -0,08 | 0,41 | -0,07 | -0,13 | 0,04 |
| Protein | 0,22 | **-0,44** | -0,13 | -0,35 | **0,65** | 0,18 | 0,18 | 0,32 | -0,14 |
| Lactose | -0,28 | **0,47** | 0,11 | -0,24 | 0,21 | 0,03 | 0,72 | -0,25 | 0,00 |
| SCC | 0,31 | 0,15 | 0,40 | **0,56** | 0,46 | -0,44 | 0,05 | 0,05 | 0,00 |
| r | **0,48** | 0,19 | -0,12 | -0,27 | 0,17 | 0,06 | -0,19 | -0,39 | 0,65 |
| k20 | **0,39** | 0,34 | 0,13 | -0,17 | -0,35 | 0,02 | 0,19 | 0,71 | 0,13 |
| A30 | **-0,47** | -0,28 | 0,02 | 0,21 | 0,06 | -0,03 | 0,16 | 0,30 | 0,73 |
| pH | 0,36 | -0,22 | -0,20 | **0,52** | -0,20 | 0,47 | 0,47 | -0,17 | 0,01 |

# PC score calculation

$$\mathbf{Z} = \begin{bmatrix} z_{11} & z_{12} & \cdots\cdots & z_{1p} \\ z_{21} & z_{22} & \cdots\cdots & z_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ z_{N1} & z_{N2} & \cdots\cdots & z_{Np} \end{bmatrix}_{(Nxp)} \qquad \mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots\cdots & a_{1p} \\ a_{21} & a_{22} & \cdots\cdots & a_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ a_{p1} & a_{p2} & \cdots\cdots & a_{pp} \end{bmatrix}_{(Nxp)}$$

$$\mathbf{Y=ZA}$$

$$\mathbf{Y} = \begin{bmatrix} Y_{11} & Y_{12} & \cdots\cdots & Y_{1p} \\ Y_{21} & Y_{22} & \cdots\cdots & Y_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ Y_{N1} & Y_{N2} & \cdots\cdots & Y_{Np} \end{bmatrix}_{(Nxp)}$$

# PCA and animal science

- ✓ **Morphology**
- ✓ **Product quality**
- ✓ **Feed analysis**
- ✓ **Management**
- ✓ **Heat stress tolerance**
- ✓ **Animal breeding and genomics**

# Use of PCA for studying lactation curve shape

J. Dairy Sci. 89:3188–3194


## Use of Multivariate Analysis to Extract Latent Variables Related to Level of Production and Lactation Persistency in Dairy Cattle

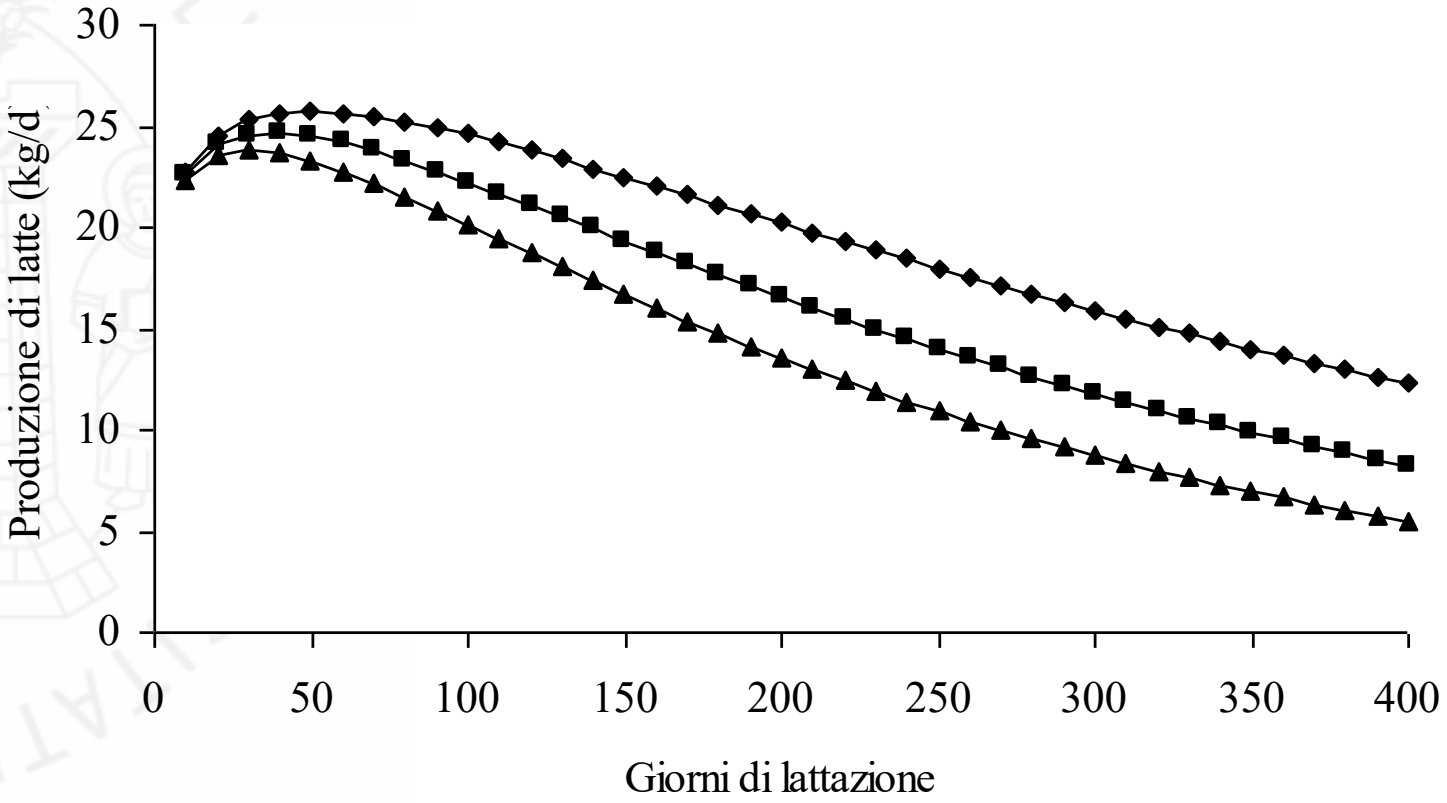
**N. P. P. Macciotta,*[1] D. Vicario,† and A. Cappio-Borlino***
*Dipartimento di Scienze Zootecniche, Università di Sassari, Via De Nicola 9, 07100 Sassari, Italy
†Italian Association of Simmental Breeders, Via Nievo 19, 33100 Udine, Italy

**ABSTRACT**

Multivariate factor analysis and principal component analysis were used to decompose the correlation matrix of test-day milk yields of 48,374 lactations of 21,721 Italian Simmental cows. Two common latent factors related to level of production in early lactation and

curve in the second part of lactation have been widely investigated in dairy cattle. Persistency of lactation, that is, the ability of a cow to maintain a constant yield during lactation (Gengler, 1996), has an economic value of about 3.4% of that for the total lactation yield (Dekkers et al., 1998). Cows with daily yield distributed

# Use of PCA for studying lactation curve shape



# Use of PCA for studying lactation curve shape

| Cow | DIM | Milk |
|---|---|---|
| 1 | 15 | 35,4 |
| 1 | 48 | 42,3 |
| 1 | 75 | 37 |
| …. | …… | …. |
| …. | …… | …. |
| 15 | 7 | 29,4 |
| 15 | 40 | 45 |
| 15 | 70 | 39 |
| …. | ….. | ….. |
| …. | ….. | ….. |
| …. | ….. | ….. |

***Daily milk yields along Lactation are usually regarded as repeated measurements in an univariate appraoch***

# Use of PCA for studying lactation curve shape

| Cow | Milk 1 | Milk 2 | Milk 3 | ….. | ….. | …… | …… | Milk 7 |
|---|---|---|---|---|---|---|---|---|
| 1 | 35,4 | 42,3 | 37 | ….. | ….. | …… | …… | 19, 2 |
| 2 | …… | …… | …… | …… | …… | …… | …… | …… |
| 3 | …… | …… | …… | …… | …… | …… | …… | …… |
| ….. | …… | …… | …… | …… | …… | …… | ….. | ……. |
| 15 | 29,4 | 45 | 39 | …… | …… | …… | ….. | 16 |

***But they can be considered under a multivariate perspective***

# PCA on over 48,000 lactation with 7 test day records each (p=7)

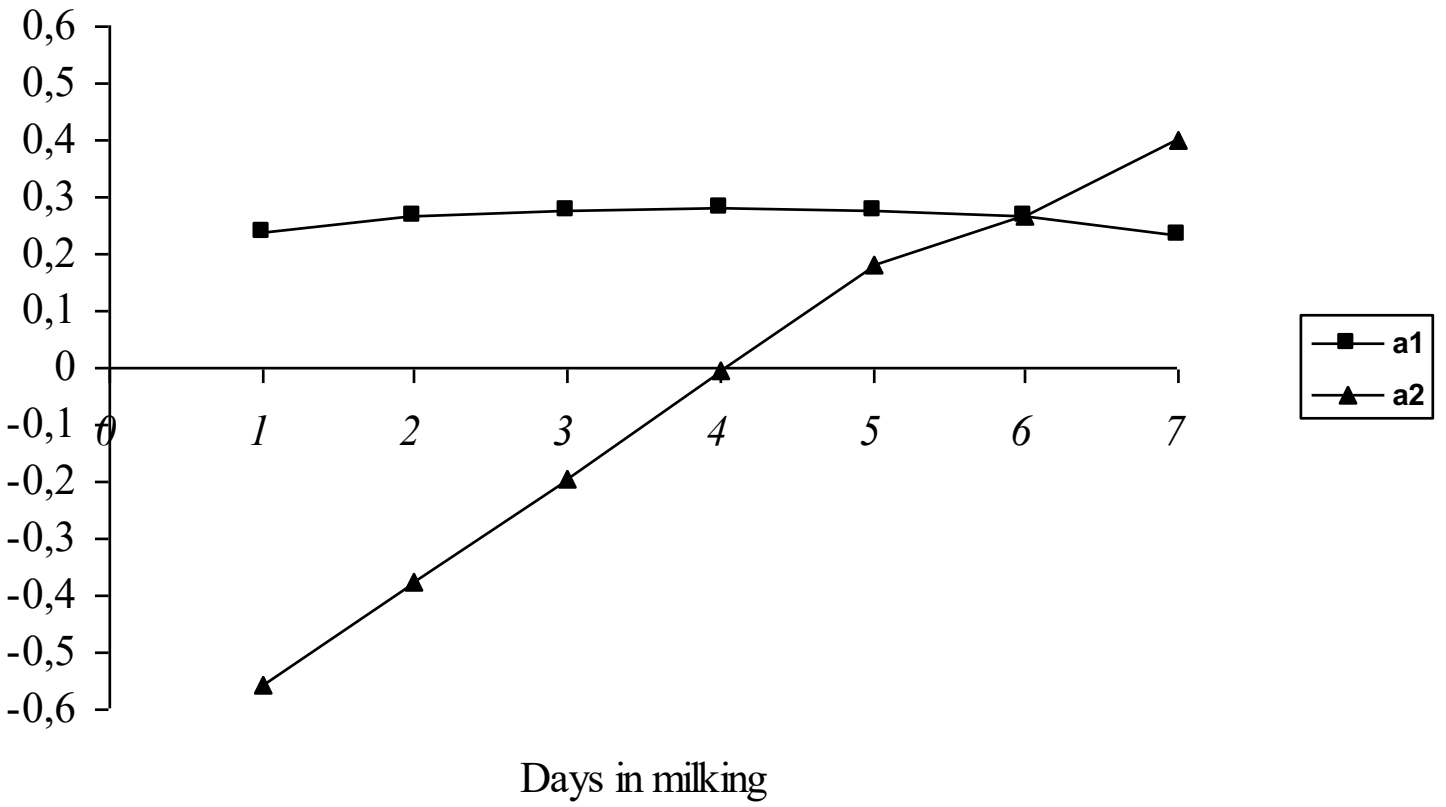
**Table 3.** Leading eigenvectors and associated eigen values of the correlation matrix of milk tests estimated by principal component analysis

| | Eigenvectors[2] | |
|---|---|---|
| Variable[1] | PC1 | PC2 |
| MILK1 | 0.345 | –0.461 |
| MILK2 | 0.396 | –0.308 |
| MILK3 | 0.411 | –0.181 |
| MILK4 | 0.415 | –0.046 |
| MILK5 | 0.409 | 0.123 |
| MILK6 | 0.383 | 0.360 |
| MILK7 | 0.261 | 0.717 |
| Eigen values (%) | 73 | 14 |

[1]MILK1 to MILK 7 = milk tests recorded at different DIM (on average 17, 60, 103, 145, 189, 232, 267, respectively).

[2]PC1 = principal component related to the average lactation yield; PC2 = principal component related to lactation persistency.

# Eigenvector plot



# Animals can be grouped according to PC scores

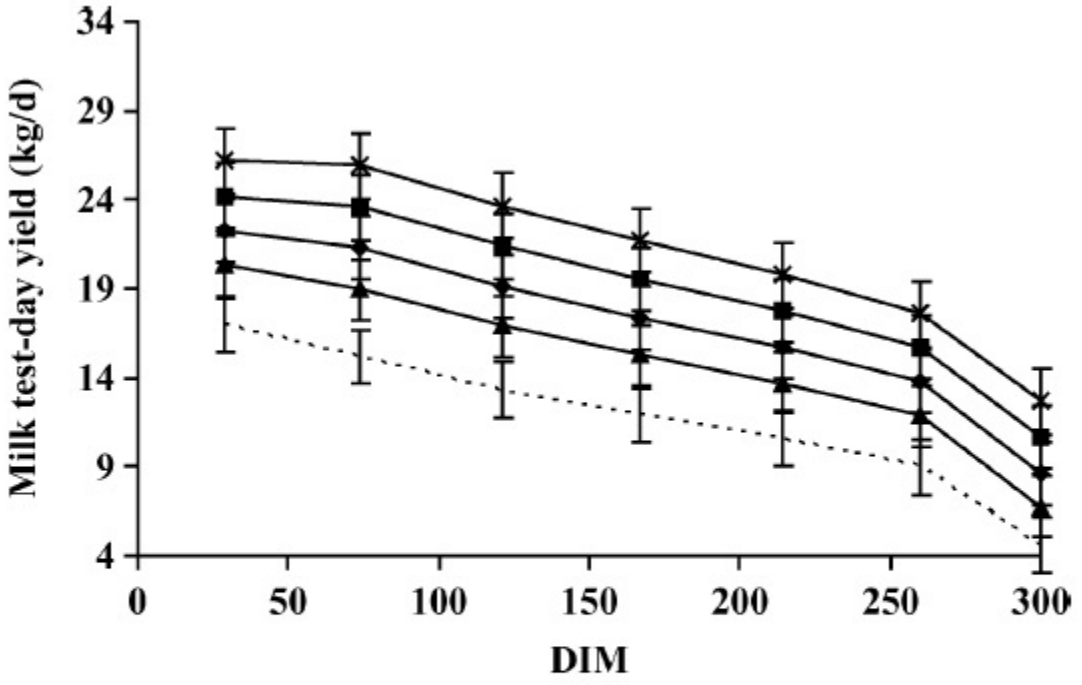



**Figure 3.** Average lactation curves of groups of cows of different PC1 score classes (· · · · = –2 to –1; ▲ = –1 to 0; ◆ = 0 to 1; ■ = 1 to 2; ● = >2). PC1 = principal component related to level of production for the whole lactation. Points are plotted for the average day in milk on each test day. Vertical bars represent the standard errors of the mean.

# Animals can be grouped according to PC scores

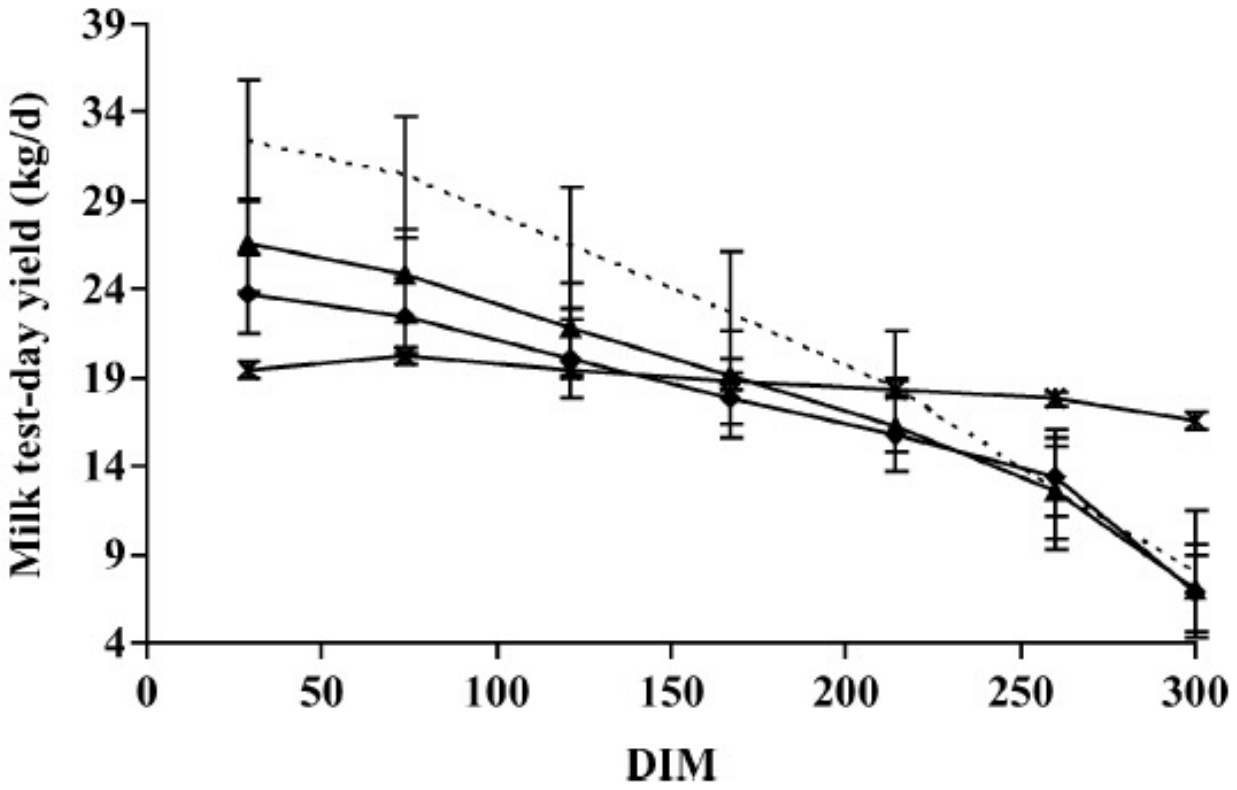



**Figure 4.** Average lactation curves of groups of cows of different PC2 score classes (· · · · = –2 to –1; ▲ = –1 to 0; ◆ = 0 to 1; ■ = 1 to 2; ● = >2). PC2 = principal component related to lactation persistency. Points are plotted for the average day in milk on each test day. Vertical bars represent the standard errors of the mean.

# *Use of PCA for studying lactation curve shape*

- ✓ Reduction of system dimensions (from 7 to 2 variables)
- ✓ Relevant amount of variance explained (87% with two principal components)
- ✓ Variables with a defined meaning (PC1=average level of production, PC2=lactation curve shape)
- ✓ PC scores able to group animals

# *Some R code for using PCA on lactation curve data*

- ✓ **Milk yield TD records (8 per animal) of 282 Italian Holstein cows**
- ✓ **Data reading and calculation of S and R**

```
data=read.table("……/dairy_cows.csv", sep=";",
header=TRUE)

cows=data[ ,3:10]

S=cov(cows)
S
R=cor(cows)
R
```

# R

| | latte1 | latte2 | latte3 | latte4 | latte5 | latte6 | latte7 | latte8 |
|---|---|---|---|---|---|---|---|---|
| latte1 | 1.00000 | 0.72719 | 0.62818 | 0.58538 | 0.52913 | 0.45254 | 0.33354 | 0.20025 |
| latte2 | 0.72719 | 1.00000 | 0.84702 | 0.74284 | 0.72049 | 0.62749 | 0.45795 | 0.31355 |
| latte3 | 0.62818 | 0.84702 | 1.00000 | 0.84072 | 0.79685 | 0.71108 | 0.57151 | 0.44148 |
| latte4 | 0.58538 | 0.74284 | 0.84072 | 1.00000 | 0.85651 | 0.75159 | 0.62385 | 0.51051 |
| latte5 | 0.52913 | 0.72049 | 0.79685 | 0.85651 | 1.00000 | 0.84531 | 0.71791 | 0.62356 |
| latte6 | 0.45254 | 0.62749 | 0.71108 | 0.75159 | 0.84531 | 1.00000 | 0.82266 | 0.73125 |
| latte7 | 0.33354 | 0.45795 | 0.57151 | 0.62385 | 0.71791 | 0.82266 | 1.00000 | 0.82216 |
| latte8 | 0.20025 | 0.31355 | 0.44148 | 0.51051 | 0.62356 | 0.73125 | 0.82216 | 1.00000 |

# PCA «by hand»

## ✓ Calculation of eigenvalues

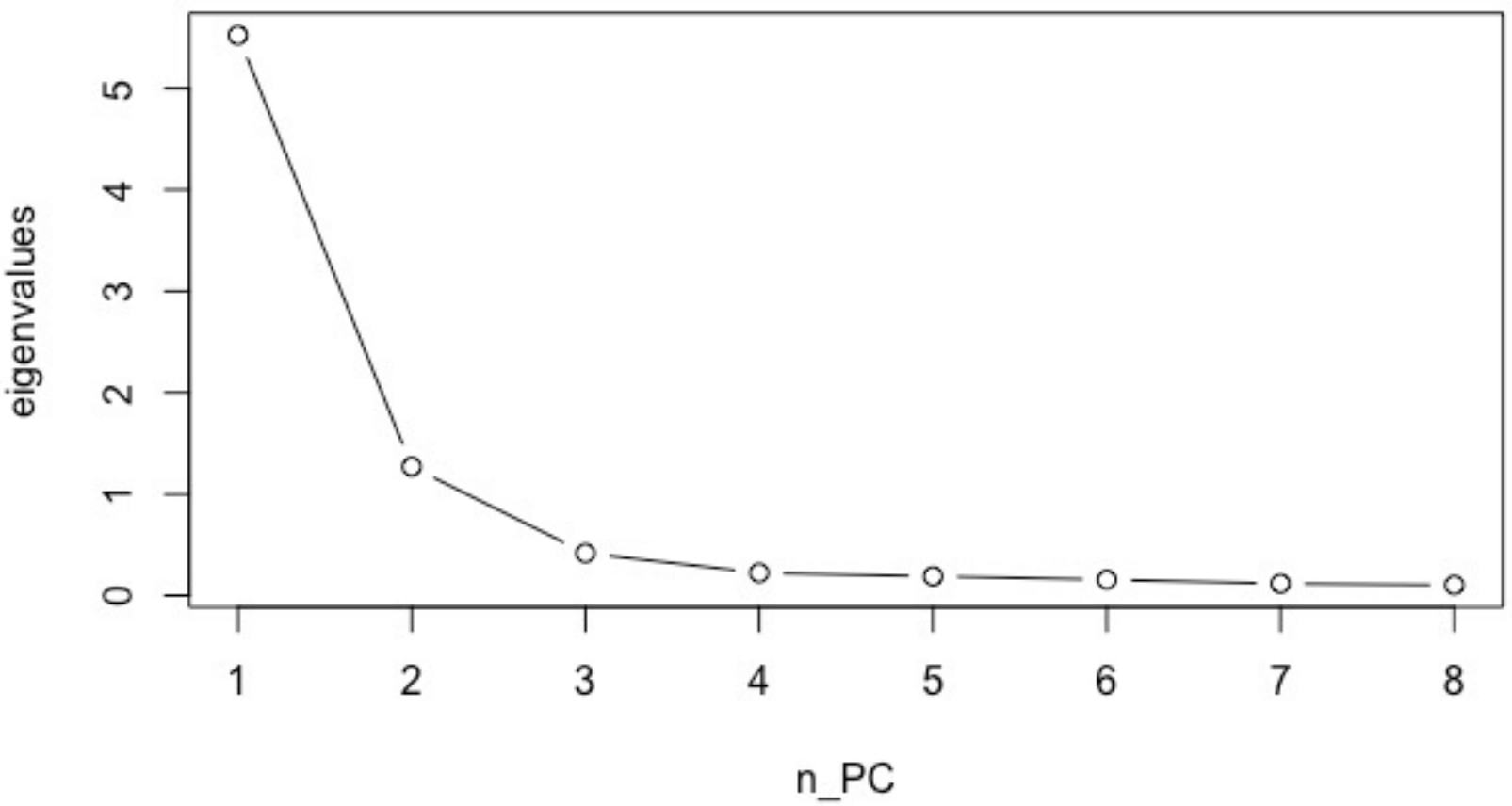
```
#eigenvalues calculation

eigenvalues=eigen(cor(vacche))$values
eigenvalues

var_tot=sum(diag(R))
var_tot

proportion=eigenvalues/var_tot
proportion

n_PC=c(1:ncol(cows))
plot(n_PC,eigenvalues, type="b")
```

# Eigenvalues

- ✓ **Eigenvalues and proportion of explained variance**

```
[1] 5.5236612 1.2696361 0.4177920 0.2228845 0.1891539 0.1559161 0.1176913
[8] 0.1032648

[1] 0.69045765 0.15870452 0.05222400 0.02786056 0.02364424 0.01948951 0.0147114
[8] 0.01290811
```

# Eigenvalue plot



# PCA «by hand»

- ✓ **Calculation of eigenvectors**

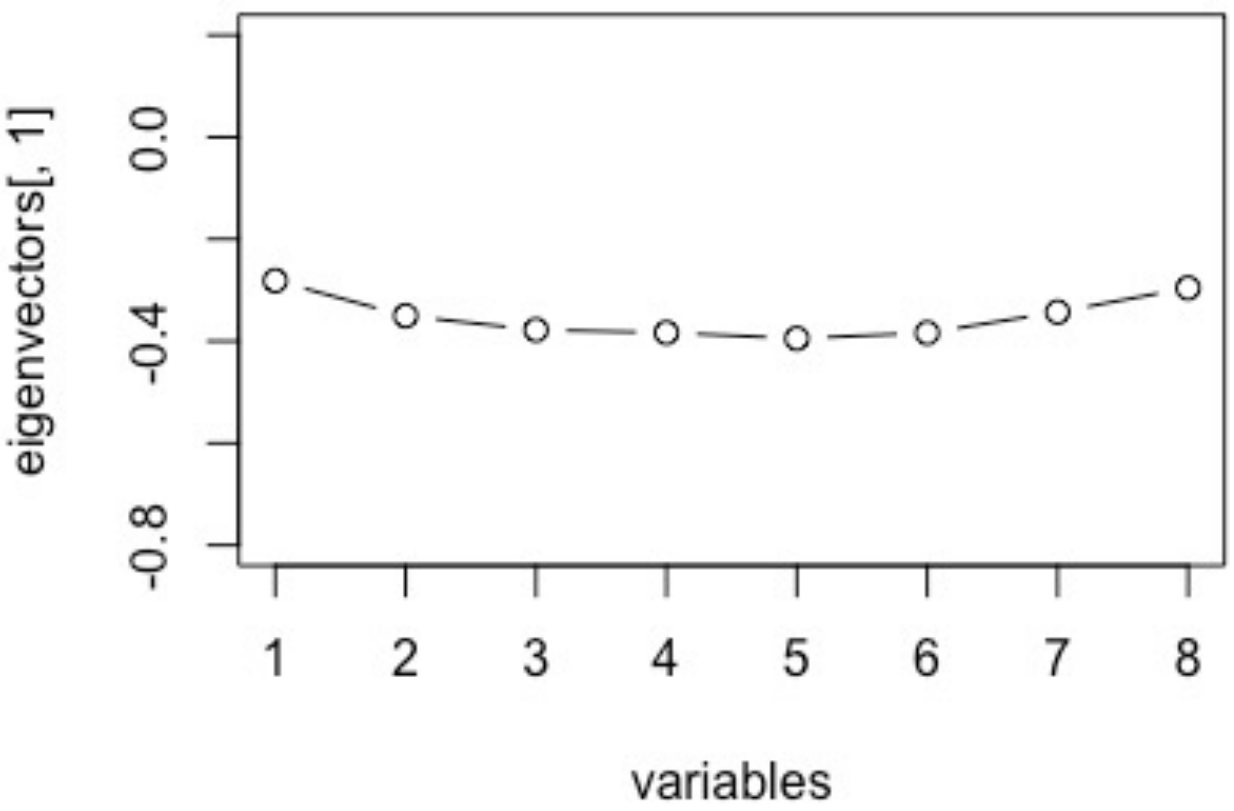
```
#eigenvector calculation
eigenvectors=eigen(cor(vacche))$vectors
eigenvectors

variables=c(1:8)

plot(variables,eigenvectors[,1], type="b", ylim=range(-
0.8:0.8))
plot(variables,eigenvectors[,2], type="b", ylim=range(-
0.5:0.5))
```

# Eigenvalues

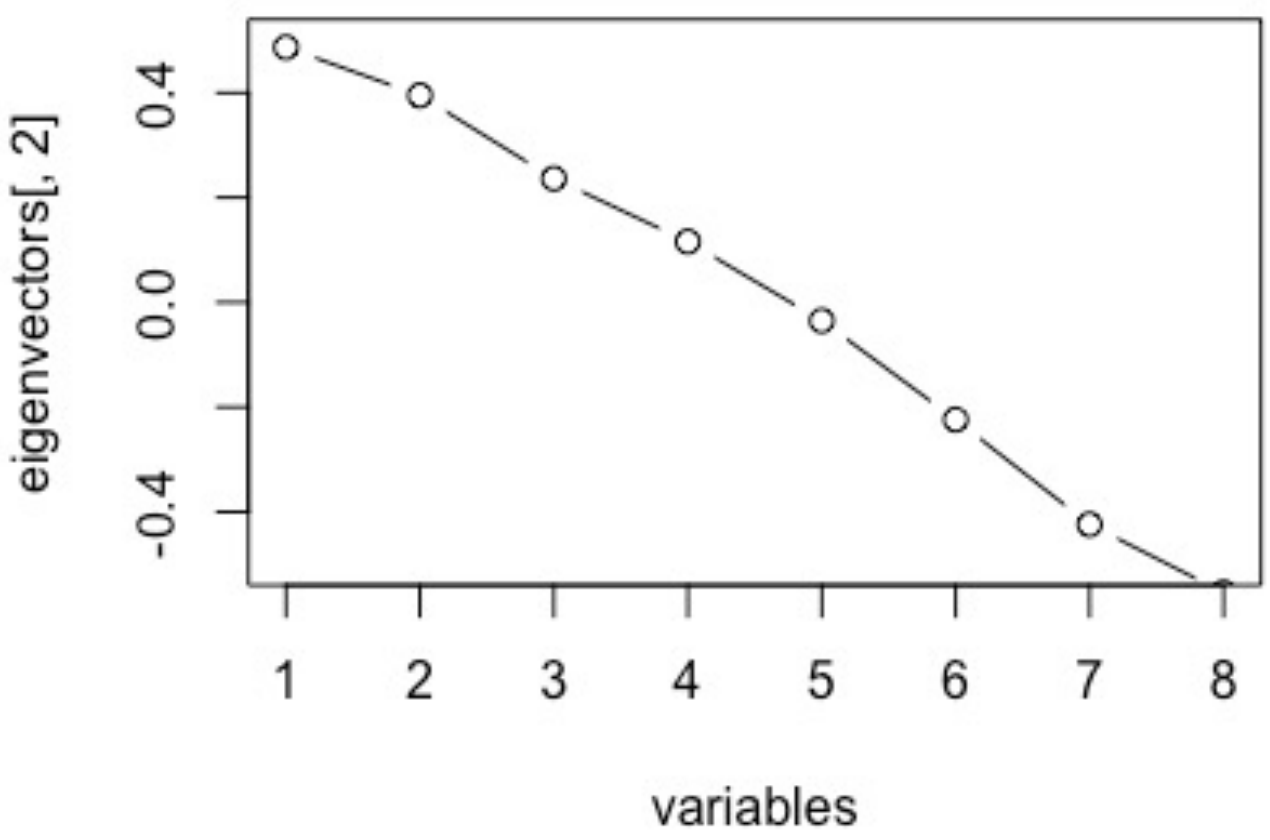
- ✓ **Eigenvectors**

| [,1] | [,2] | [,3] | [,4] | [,5] | [,6] | [,7] | [,8] |
|---|---|---|---|---|---|---|---|
| -0.28137 | 0.486666 | 0.758447 | -0.291533 | -0.037271 | -0.0041882 | -0.102272 | 0.109031 |
| -0.35030 | 0.394930 | -0.039030 | 0.599901 | 0.115930 | 0.2270182 | 0.397677 | 0.369853 |
| -0.37798 | 0.236208 | -0.300253 | 0.355379 | -0.283832 | -0.2671467 | -0.463276 | 0.467263 |
| -0.38338 | 0.115917 | -0.352569 | -0.511633 | -0.360909 | -0.2292060 | 0.037815 | -0.518932 |
| -0.39479 | -0.035076 | -0.268064 | -0.349776 | 0.191691 | 0.3782594 | 0.419748 | 0.541010 |
| -0.38380 | -0.223869 | -0.026450 | -0.050489 | 0.618680 | 0.1865963 | -0.563904 | -0.252502 |
| -0.34303 | -0.423934 | 0.236480 | 0.125995 | 0.199345 | -0.6850297 | 0.343321 | 0.062781 |
| -0.29570 | -0.553974 | 0.283366 | 0.162162 | -0.561343 | 0.4211631 | -0.057389 | -0.057494 |

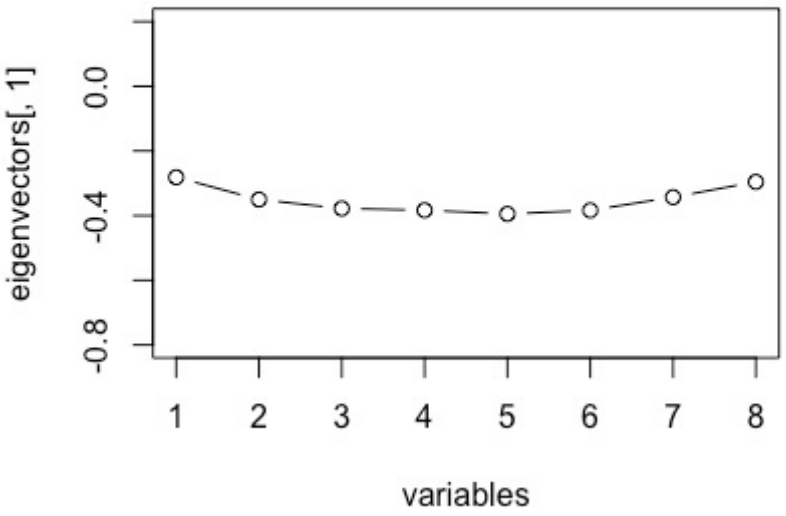
# Eigenvector plot PC1



# Eigenvector plot PC2

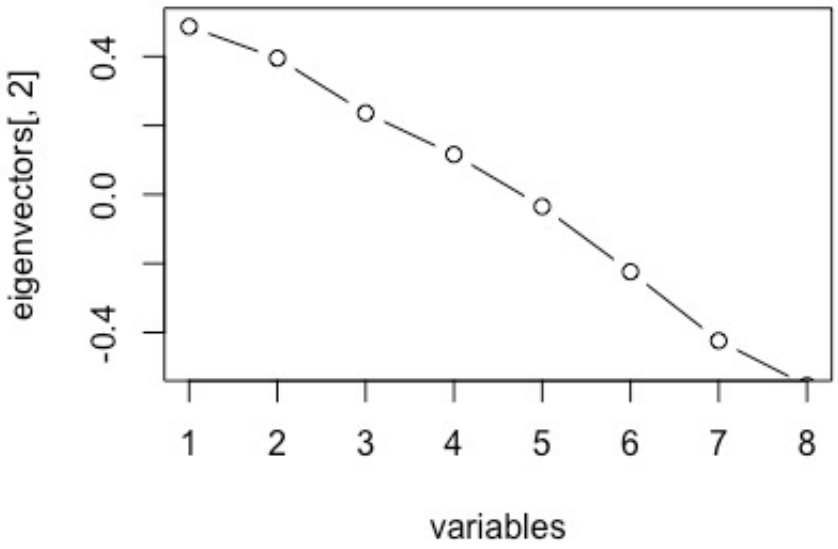


# Eigenvector plot




**Eigenvector PC1**

**Eigenvector PC2**



# PC «by hand»

## ✓Calculation of PC scores

```
#Standardization of data Z
p=8
N=482
st_dev=as.matrix(sapply(cows,sd))
mean=as.matrix(sapply(cows,mean))
Z=matrix(0,nrow=N,ncol=p)
for (i in 1:N) {
  for (j in 1:p) { Z[i,j]=(cows[i,j]-mean[j,1])/st_dev[j,1]
                 }
         }

#PC scores calculation
PC_scores=Z%*%eigenvectors
PC_scores
```

# Scores

| | [,1] | [,2] | [,3] | [,4] | [,5] |
|---|---|---|---|---|---|
| [1,] | -3.314268987 | -1.427369222 | 0.8411749937 | -0.1851525221 | -0.007402420 |
| [2,] | -3.351244799 | 1.420043466 | 0.1909905660 | -0.2737439602 | 0.358970041 |
| [3,] | -2.603960860 | 1.673484840 | 0.2473763074 | -0.4420107998 | -0.248361015 |
| [4,] | 0.259124494 | 1.622038769 | -0.8071538573 | 0.0521694772 | 0.136637319 |
| [5,] | -2.499730091 | 0.970448786 | -2.0242262774 | -0.4626567705 | 0.271432725 |
| [6,] | -4.222462564 | -0.888561740 | -0.7107824309 | -0.2315899981 | 0.172094281 |
| [7,] | -1.270731925 | 0.928956514 | -0.9297448367 | 0.1924907490 | 0.420002944 |

# Scores can be treated as new variables

```
#evaluation of the effect of parity on lactation persistency
parity=data[,2]
parity=factor(parity)
persistency=PC_scores[ ,2]
analysis=lm(persistency~parity)
test=anova(analysis)
test
summary(analysis)

job=aov(analysis)
TukeyHSD(job)
```

# Scores can be treated as new variables

Analysis of Variance
Table Response: persistenza

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| parity | 6 | 117.8 | 19.6326 | 18.92 | < 2.2e-16 *** |
| Residuals | 475 | 492.9 | 1.0377 | | |

--- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Call: lm(formula = persistenza ~ parity)
Residuals:

| Min | 1Q | Median | 3Q | Max |
|---|---|---|---|---|
| -3.2355 | -0.6350 | -0.0082 | 0.6241 | 2.9913 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | -0.70527 | 0.08402 | -8.394 | 5.43e-16 *** |
| parity2 | 0.80097 | 0.12030 | 6.658 | 7.67e-11 *** |
| parity3 | 1.21420 | 0.12651 | 9.598 | < 2e-16 *** |
| parity4 | 1.02135 | 0.17661 | 5.783 | 1.33e-08 *** |
| parity5 | 1.28149 | 0.23764 | 5.393 | 1.10e-07 *** |

# Scores can be treated as new variables

Tukey multiple comparisons of means 95% family-wise confidence level
Fit: aov(formula = analisi) $parity

| | diff | lwr | upr p | adj |
|---|---|---|---|---|
| 2-1 | 0.80097418 | 0.44477497 | 1.1571734 | 0.0000000 |
| 3-1 | 1.21420190 | 0.83960437 | 1.5887994 | 0.0000000 |
| 4-1 | 1.02135283 | 0.49840472 | 1.5443009 | 0.0000003 |
| 5-1 | 1.28149040 | 0.57783339 | 1.9851474 | 0.0000023 |
| 6-1 | 1.06374721 | 0.22009208 | 1.9074023 | 0.0039611 |
| 7-1 | 1.23605101 | -1.79049225 | 4.2625943 | 0.8904734 |
| 3-2 | 0.41322772 | 0.03452217 | 0.7919333 | 0.0222632 |
| 4-2 | 0.22037865 | -0.30551993 | 0.7462772 | 0.8778115 |
| 5-2 | 0.48051622 | -0.22533630 | 1.1863687 | 0.4059917 |
| 6-2 | 0.26277302 | -0.58271414 | 1.1082602 | 0.9692545 |
| 7-2 | 0.43507683 | -2.59197763 | 3.4621313 | 0.9995433 |
| 4-3 | -0.19284907 | -0.73137924 | 0.3456811 | 0.9391911 |
| 5-3 | 0.06728850 | -0.64802486 | 0.7826019 | 0.9999616 |
| 6-3 | -0.15045470 | -1.00385611 | 0.7029467 | 0.9985330 |
| 7-3 | 0.02184911 | -3.00742542 | 3.0511236 | 1.0000000 |
| 5-4 | 0.26013757 | -0.54287216 | 1.0631473 | 0.9622937 |
| 6-4 | 0.04239438 | -0.88574645 | 0.9705352 | 0.9999995 |

# Genomic selection

- Use of medium (50K) and high (800K) density SNP map for predicting the genetic merit of individuals
- Two step – One step approach
- Animals genotyped and phenotyped (reference), only genotyped, or only phenotyped (females)

# Genomic selection

- ✓ Data matrix with particular structure
- ✓ Only three numbers (0,1,2)
- ✓ Columns >>Rows

$$\mathbf{Z} = \begin{bmatrix} 0 & -1 & \ldots\ldots\ldots & 1 \\ 1 & 0 & \ldots\ldots\ldots & 1 \\ \vdots & \vdots & \vdots & \vdots \\ -1 & 0 & \ldots\ldots\ldots & 0 \end{bmatrix}_{(\mathrm{Nxp})}$$

# The problem of dimensionality

- ✓ Information contained in SNP genotypes could be compressed using dimension reduction techniques
- ✓ Multivariate techniques as principal component analysis or partial least squares regression have been successfully used in GS (Solberg et al., 2009; Pintus et al., 2102)

# Compress the information contained in a large number of SNP in a small number of Principal Components

| Breed | n. | SNP | PC | Explained variance |
|---|---|---|---|---|
| Holstein | 2093 | 40,658 | 2,564 | 80% |
| Brown | 749 | 37,254 | 2,257 | 80% |
| Simmental | 479 | 40,179 | 2,476 | 70% |

J. Dairy Sci. 95:3390–3400
http://dx.doi.org/10.3168/jds.2011-4274


# Prediction of genomic breeding values for dairy traits in Italian Brown and Simmental bulls using a principal component approach

M. A. Pintus,* G. Gaspa,* E. L. Nicolazzi,† D. Vicario,‡ A. Rossoni,§ P. Ajmone-Marsan,† A. Nardone,# C. Dimauro,* and N. P. P. Macciotta*[1]

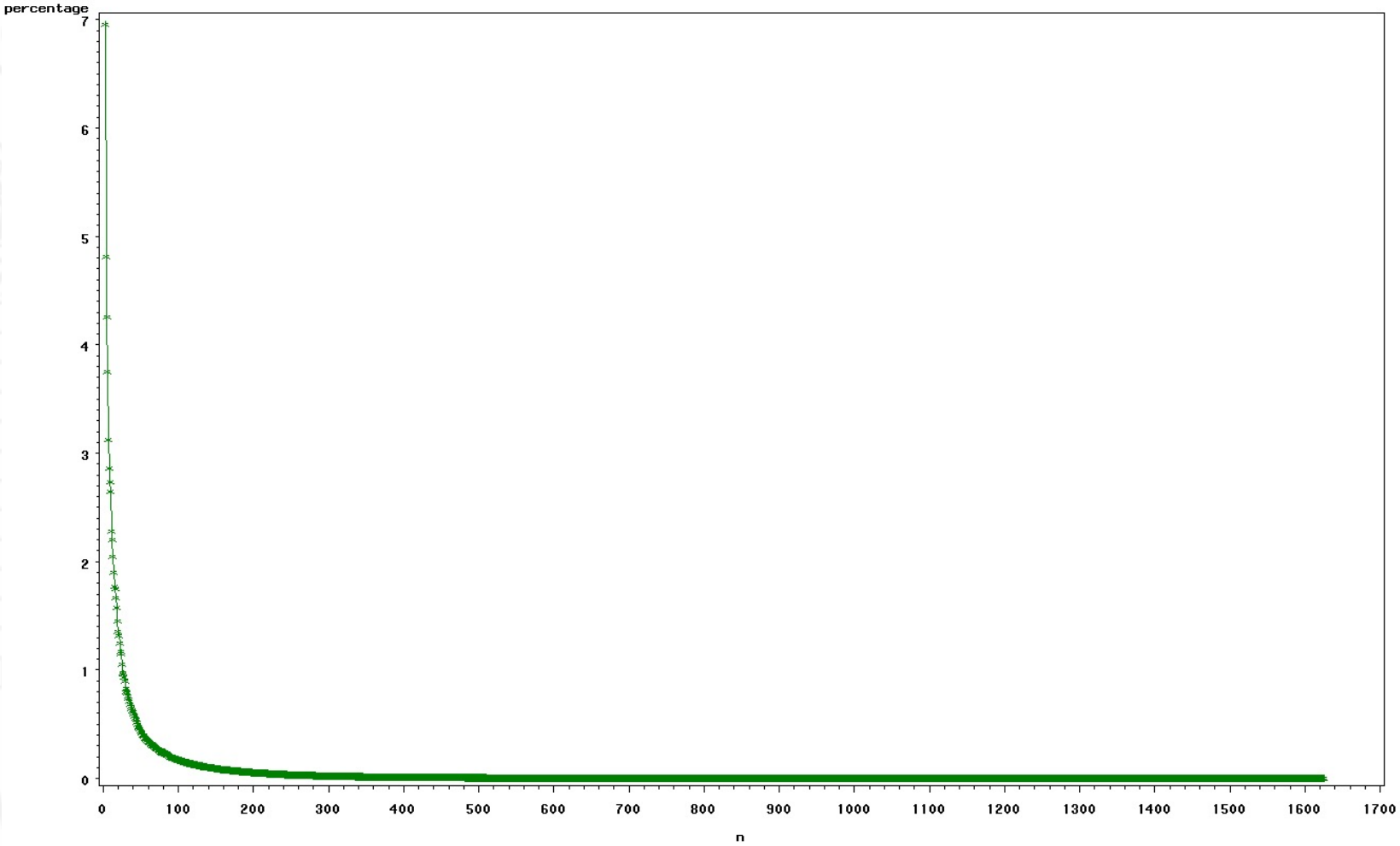
## Squared correlations between GEBV and polygenic EBV in prediction bulls (80:20)

| Trait | SNP BLUP | PC BLUP | Bayes A | PI |
|---|---|---|---|---|
| Milk yield | 36.7 | 35.3 | 35.7 | 33.1 |
| Fat yield | 31.2 | 30.0 | 30.3 | 28.8 |
| Protein Yield | 33.0 | 30.6 | 31.0 | 30.5 |
| SCC | 20.3 | 20.5 | 20.6 | 20.6 |
| Fat percentage | 12.7 | 14.9 | 14.1 | 15.9 |
| Protein percentage | 17.9 | 16.5 | 17.1 | 16.9 |

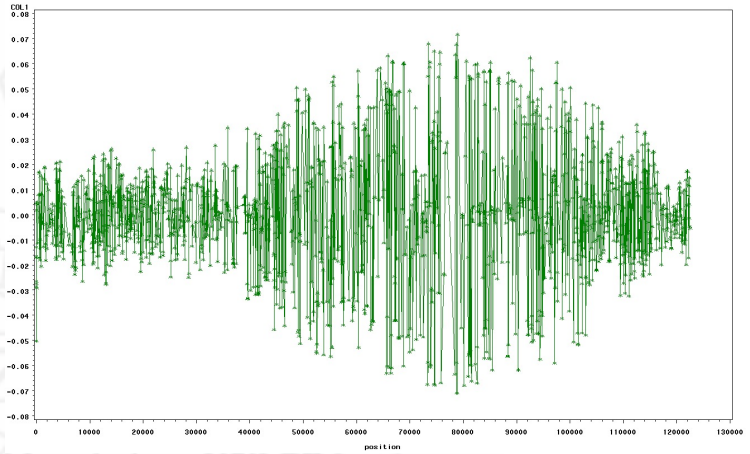
# Some considerations

- ✓ **Relevant reduction of predictor dimensionality (>90%)**
- ✓ **PCA based method gave similar performance compared to the use of all predictors**
- ✓ **Differences between traits**
- ✓ **Differences between breeds**

# Eigenvalues plot

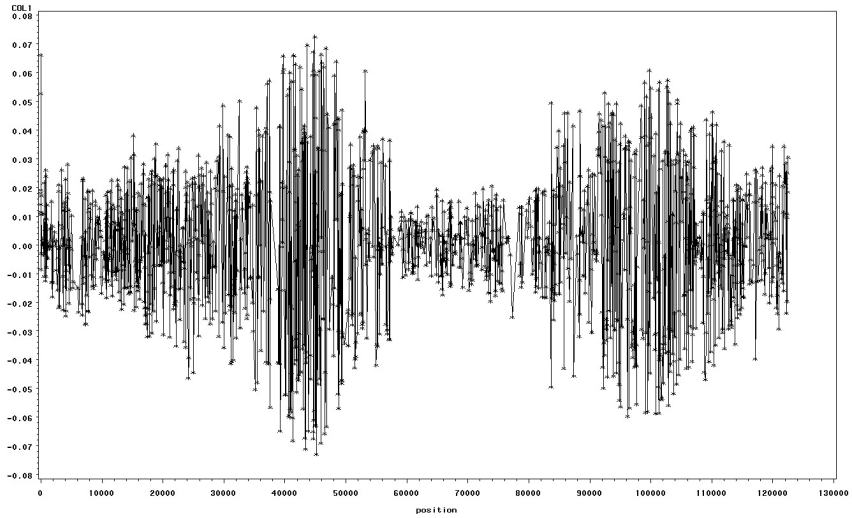


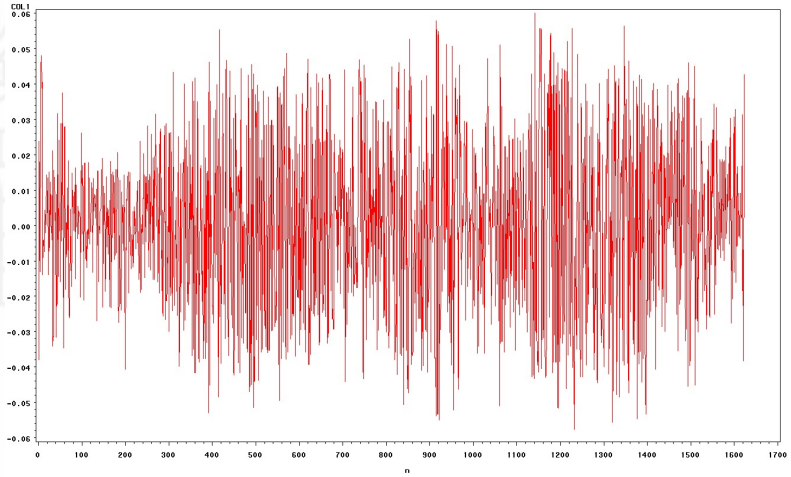
# Difficult to intepret eigenvectors




PC1

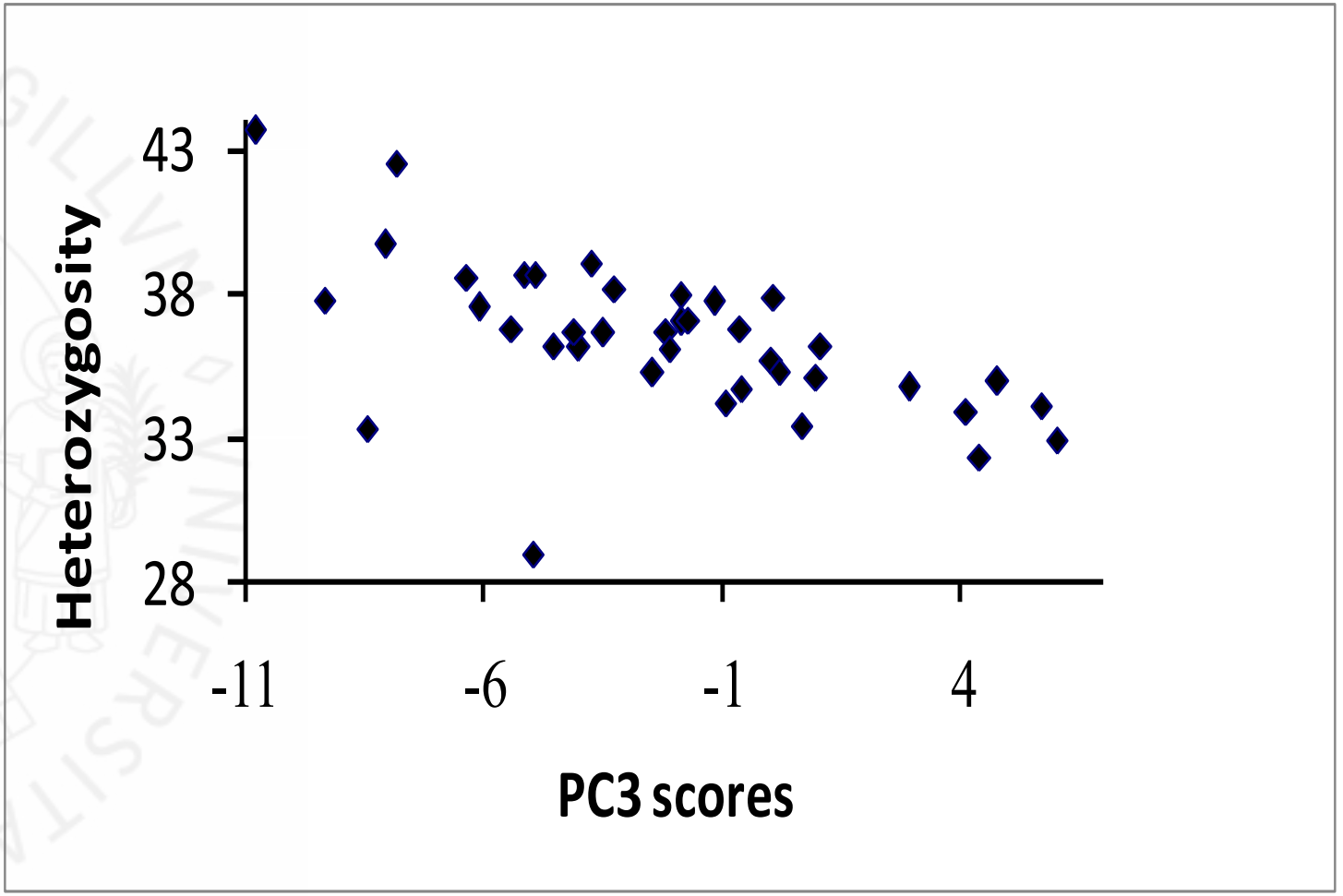
PC2




PC3

# Infer PC meaning from relationships with other variables




($r_{PC3,het}$=-0.43)

# PC could be used to discriminate generations



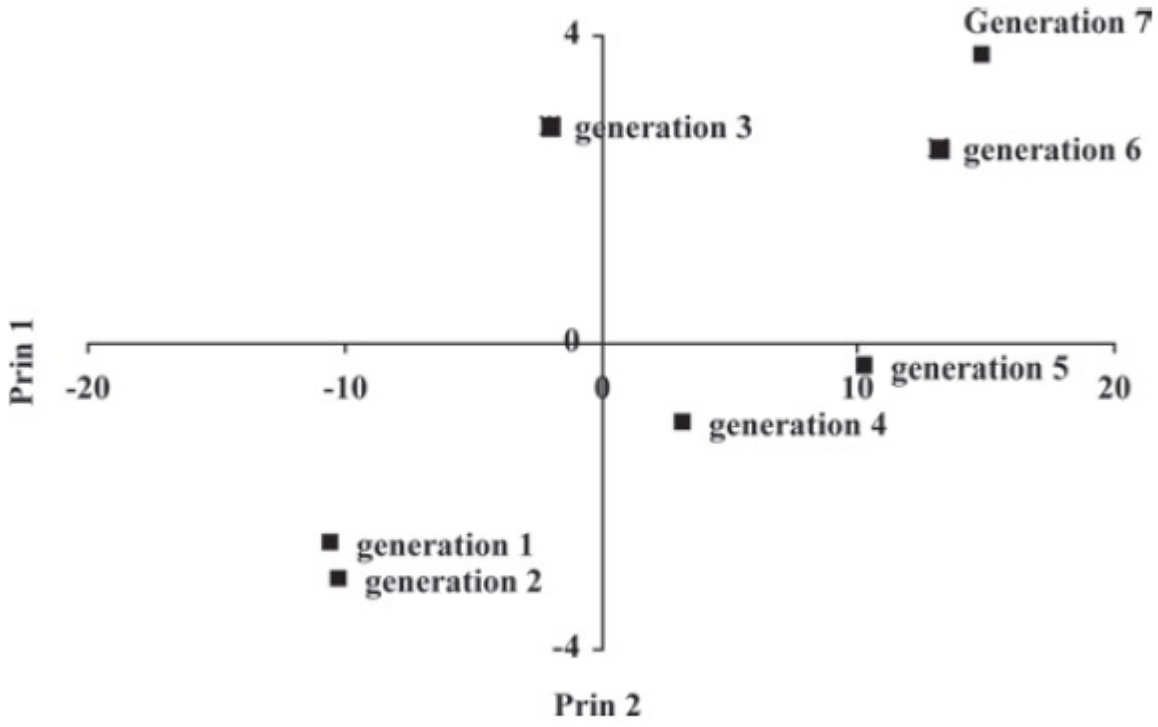
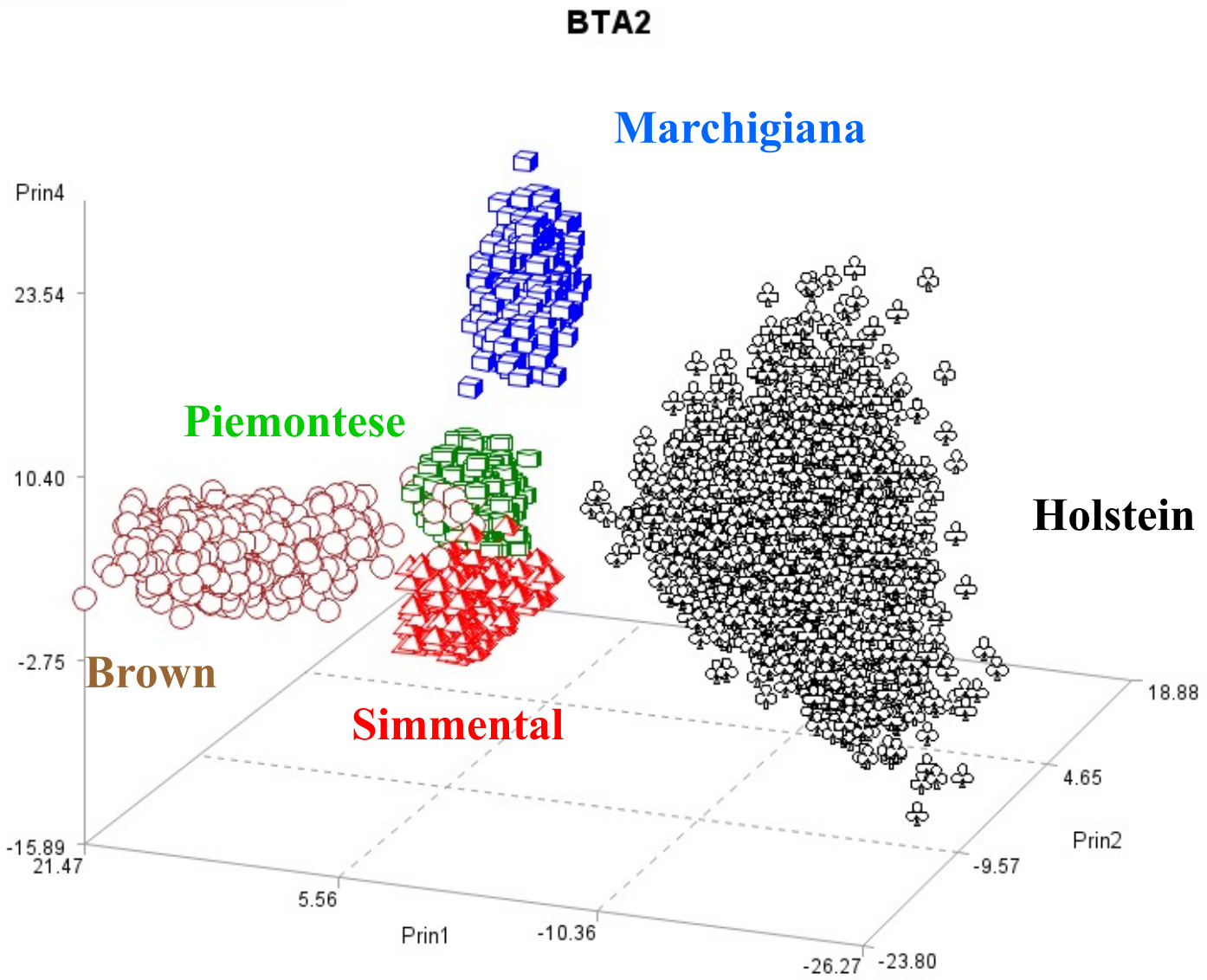


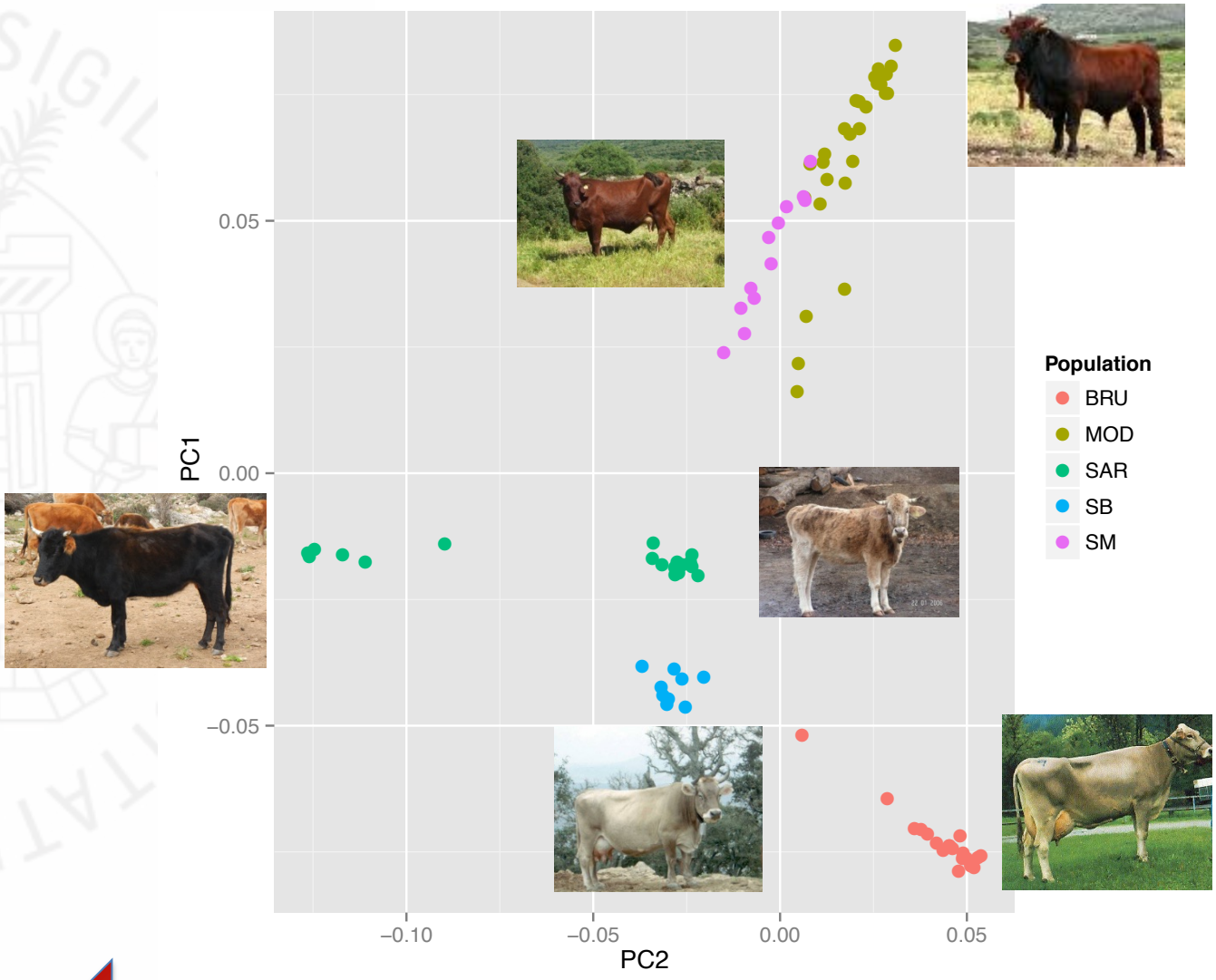

**Figure 6.** Plot of the average scores of the first 2 principal components (Prin) for 7 generations.

**(Macciotta et al., JDS 2010)**

# Or to discriminate between breeds



# Main PC have often a geographical meaning



# *Use of PCA for reducing prediction dimensionality in genomic data*

- ✓ Reduction of system dimensions (about 95%)
- ✓ Relevant amount of variance explained (70-80% with two principal components)
- ✓ Variables able to distinguish animals of different breeds/families

# R has some functions for performing PCA

#**prcomp**

```
PCA=prcomp(cows, cor=TRUE)
PCA
summary(PCA)
biplot(PCA)
```

#**princomp**

```
PCA=princomp(cows, cor=TRUE)
PCA
summary(PCA)
loadings(PCA)
```

# *Singular value decomposition SVD*

SVD decomposes a matrix **M** into the product of two unitary matrices (**U, V**) and a rectangular diagonal matrix **Σ** with single values on the diagonal

$$\mathbf{M}_{(nxm)} = \mathbf{U}_{(nxn)}\, \mathbf{\Sigma}_{(nxm)}\, \mathbf{V}'_{(mxm)}$$

- If SVD is applied to correlation or covariance matrix, it is equivalent to PCA

- More efficient in term of computational power P

# *Use of PCA for farm management analysis*

J. Dairy Sci. 96:3378–3387
http://dx.doi.org/10.3168/jds.2012-6256


## A multivariate and stochastic approach to identify key variables to rank dairy farms on profitability

**A. S. Atzori,*[1] L. O. Tedeschi,† and A. Cannas***
*Dipartimento di Agraria, Sezione di Scienze Zootecniche, Università di Sassari, 07100 Sassari, Italy
†Department of Animal Science, Texas A&M University, College Station 77843-2471

**ABSTRACT**

The economic efficiency of dairy farms is the main goal of farmers. The objective of this work was to use routinely available information at the dairy farm level to develop an index of profitability to rank dairy farms and to assist the decision-making process of farmers to increase the economic efficiency of the entire system. A stochastic modeling approach was used to study the relationships between inputs and profitability (i.e., income over feed cost; IOFC) of dairy cattle farms. The IOFC was calculated as: milk revenue + value of male calves + culling revenue − herd feed costs. Two databases were created. The first one was a development database, which was created from technical and economic variables collected in 135 dairy farms. The calculated, we found that 21 farms had dRI < −1 SD, 32 farms were between −1 SD and 0, 67 farms were between 0 and +1 SD, and 15 farms had dRI > +1 SD. The top 10% of the farms had a dRI greater than 0.170 €/L, whereas the bottom 10% farms had a dRI lower than 0.116 €/L. This stochastic approach allowed us to understand the relationships among the inputs of the studied dairy farms and to develop a ranking index for comparison purposes. The developed methodology may be improved by using more inputs at the dairy farm level and considering the actual cost to measure profitability.
**Key words:** principal component analysis, ranking index, decision making unit, modeling

# Results (dDB; n=135 farms)

## Mean values of main input variables and distribution

| Input variable | Mean | DS | @Risk -best fit distribution* |
|---|---|---|---|
| Equiv.MatureCow, kg/cow/Yr | 10,294 | 946 | Normal |
| Primiparous, n° | 41.6 | 25.4 | Loglogistic |
| Pluriparous, n° | 72.4 | 40.3 | Gamma |
| Drycows, n° | 19.0 | 11.2 | Invgauss |
| Replacement, n° | 52.0 | 30.1 | Gamma |
| Age1calv, months | 28.3 | 4.4 | Normal |
| Agecull, months | 68.4 | 13.6 | Invgauss |
| Days open, days | 164.4 | 26.9 | Extvalue |
| Soldmilk, lt x 1000 | 974.2 | 567.3 | Weibull |
| Fat,% | 3.8 | 0.1 | Gamma |
| Prot,% | 3.4 | 0.1 | Gamma |
| Milkprice, $ | 0.35 | 0.14 | Weibull |

# RESULTS (sDB; n=5000 farms)

| Input\PCA | HERD PROFILE | MILK QUALITY and PAYMENT | BAD MANAGEMENT | REPRODUCTION |
|---|---|---|---|---|
| Primiparous | **0.35** | 0.04 | 0.05 | -0.04 |
| Pluriparous | **0.37** | 0.03 | 0.15 | -0.03 |
| Drycows | **0.35** | 0.02 | 0.17 | 0.12 |
| Replacement | **0.36** | 0.03 | 0.13 | -0.08 |
| Calvings | **0.37** | 0.02 | 0.10 | -0.03 |
| Culled cows | **0.35** | - 0.01 | 0.12 | 0.00 |
| Sold milk | **0.38** | 0.04 | 0.05 | -0.01 |
| Fat | -0.09 | **0.41** | 0.14 | -0.18 |
| Protein | -0.09 | **0.36** | 0.23 | -0.10 |
| Milk price | 0.02 | **0.48** | -0.16 | 0.11 |
| Fat bonus | -0.09 | **0.40** | 0.15 | -0.18 |
| Prot bonus | -0.07 | **0.38** | 0.19 | -0.06 |
| Other bonus | 0.08 | **0.32** | -0.27 | 0.17 |
| Equiv. mature cow | 0.12 | 0.03 | **-0.40** | 0.22 |
| Age 1 calving | -0.11 | -0.04 | **0.38** | 0.12 |
| Age Culling | -0.11 | -0.08 | **0.40** | 0.26 |
| SCC BONUS | 0.10 | 0.22 | **-0.38** | 0.32 |
| Days Open | -0.07 | 0.06 | 0.26 | **0.79** |

uniss
UNIVERSITÀ DEGLI STUDI DI SASSARI

# RESULTS (sDB; n=5000 farms)

| EXTRACTED PRINCIPAL COMPONENT | EXPLAINED VARIANCE in the SDB |
|---|---|
| PC1 Herd profile | 37.4 % |
| PC2 Milk quality and payment | 20.2 % |
| PC3 Management | 15.6 % |
| PC4 Reproduction | 4.4 % |
| Total | 77.6 % |

IOFC ($/liter) =

0.0016*PC1 + 0.0058*PC2 - 0.0082*PC3 - 0.005*PC4

*Adj. $R^2$ = 0.72*

uniss
UNIVERSITÀ DEGLI STUDI DI SASSARI

# Some R code

```
#------------------------ Calculation of correlation matrix R-----------------------

devX=apply(X,2,sd)

Ds=diag(devX)

L=solve(Ds)

Z=D%*%L

# R= Z'Z/(n-1)
R<-1/(n-1)*t(Z)%*%Z


# ---------- R functions for calculating S o R
var(X)
cor(X)
```

# Some R code

```
#----------------Calculation of variance-covariance matrix S--------------------------------
X=matrix(c(168,184,173,176,176,72,75,58,58,68,30,29,26,26,28),ncol=3)

Xm=apply(X,2,mean)

Xm=diag(Xm)

U=matrix(1,nrow=nrow(X),ncol=ncol(X))

D=X-U%*%Xm

n=nrow(X)

#S= 1/(n-1)XD'D variance-covariance matrix calculation

S=1/(n-1)*t(D)%*%D
```