Unsupervised Multivariate Methods for “Omics” Data Analysis

Alex Sánchez

Statistics Department. University of Barcelona

Statistics and Bioinformatics Unit. VHIR.

Outline

- Introduction. Why multivariate analysis?
- Descriptive / Exploratory methods
- Dimension reduction: PCA and relatives
- Distances between objects
- Finding patterns in data.
- Where to now ...

The nature of omics data

- Omics data are diverse
  - They measure distinct characteristics
    - GC/MS spectrum, Expression, Concentration…
- Although they have aspects in common
  - Most of them are high throughput
    - Many variables (**K**) measured simultaneously
  - Relatively expensive, ethical limitations, regulations
    - Few samples (**N**) analyzed

$K >> N$

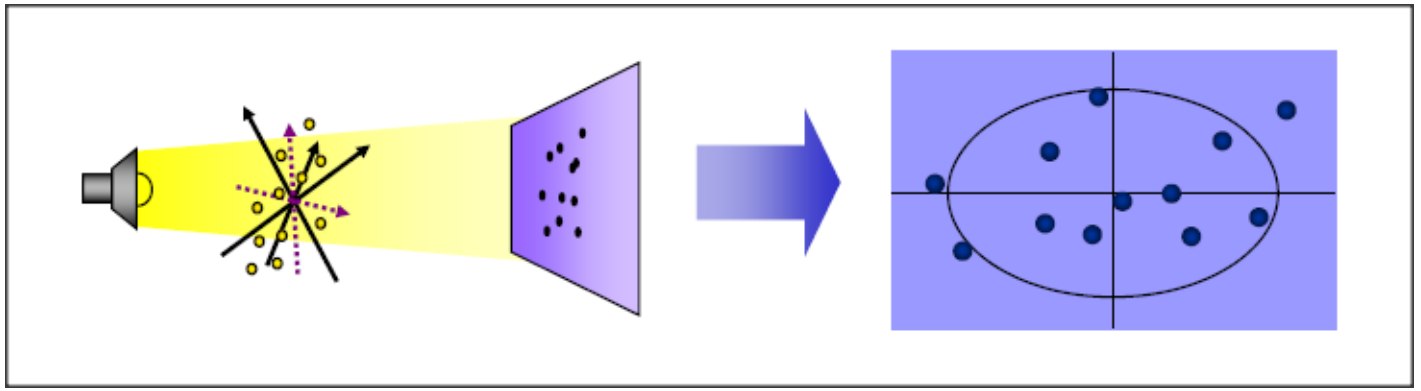
Best approach for omics data analysis?

- Classical Statistics
  - Multiple regression
  - Discriminant analyisis
  - ANOVA
- Data tables are *long* and *lean*

- Assumptions
  - Independent variables
  - More observations than variables
  - Multivariate normality
  - Interested in one dependent
  - Few missings
- DO NOT hold for many omics data

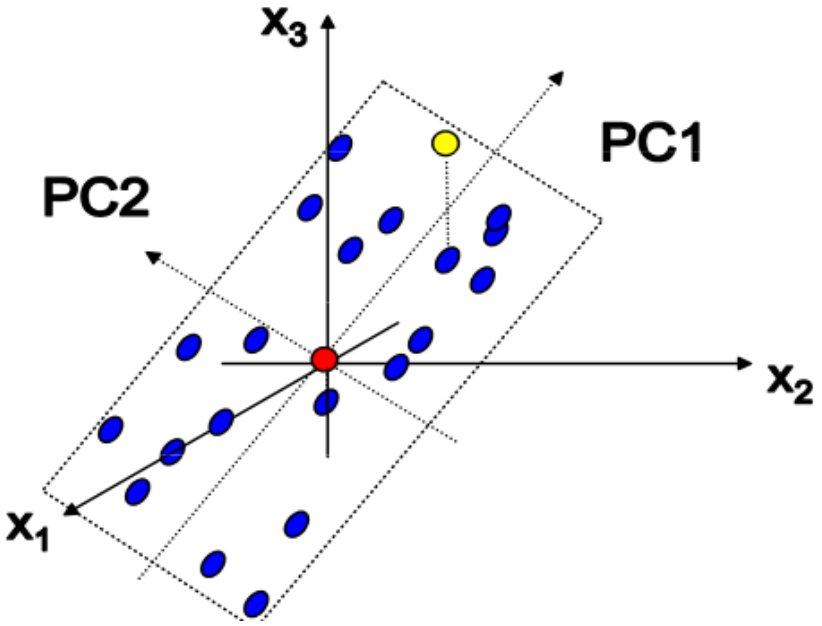
A Better Way

- Multivariate analysis by *projection methods*
  - Looks at ALL the variables together
  - Avoids loss of information
  - Finds underlying trends = “latent variables”
  - More stable models



What is a projection?

- Variables = axes in a multidimensional space
- Observations = points in multidimensional space
- By projecting points onto (hyper)planes
  - Visualization is simplified
  - Relation between them is highlighted
    - Show natural groups
    - Can help detect problems such as batch effects or outliers



- Algebraically:
  - The information in the observations is summarized as a few new (latent) variables
- Geometrically:
  - The swarm of points in a K dimensional space (K = number of variables) is approximated by a (hyper)plane and the points are projected on that plane

Approaches considered here

- Descriptive multivariate
  - Always start by looking/exploring the data
- Projection/Dimension reduction methods
  - Principal components analysis (PCA), Multidimensional scaling (MDS)
    - Look at the data by means of projection
- Find patterns in data
  - Cluster analysis

Data examination

- Start any statistical analysis by looking at the data
  - Which/How many variables,
  - Which/How many samples
  - Are there any missing values
  - Obtain simple summary statistics and plots

Data examination: numerical summaries

- The ***variance*** of a variable is a measure of the spread of its values.
  - The ***standard deviation***, square root of the variable expresses the measure of the spread in the same unit as the measurements.
- The ***covariance*** between two variables is a measure of their linear association.
- Covariance depends on the units of the variables. Use the ***correlation coefficent*** to have a unitless scale.
- ***Skewness*** is a measure of the simmetry of the variable. Vaues outside [-1, +1] suggest skewed distributions.

Data examination; plots

- Graphs are typically 2D or 3D
  - Hard to plot more than 3 variables simultaneously
    - Different alternatives exist
      - Work on reduced dimension
      - Plot variables separately (do not account for their relation)
      - Intermediate situations
- Many different methods
  - Box plots: visualize five number summary
  - Pairs plots: matrix of scatterplots (many 2D plots)
  - Star plot: Each observation: star-shaped figure with one ray per variable

Example: Exploring an expression matrix

- A study was performed using affymetrix microarrays to study the effect of DMSO on cancer cell lines
- Samples from 3 individuals were treated with
  - DMSO
  - Dioxin, that acted as control
- HGU95AV2 (A and B) arrays were performed and are avalable at GEO as GSE7765 series.

The data: Expression matrix

- Microarray data→ Expression matrices
  - Intensities obtained by scanning the array
    - Prop. to [mRNA] of genes expressed in the sample
  - Usually, gone through several preprocessing steps
    - Normalization, Logarithms, Centering/Scaling,

	DMSO013	Diox014	DMSO016	Diox018	DMSO020	Diox022
1007_s_at	15630.200	17048.800	13667.500	15138.800	10766.600	15680.800
1053_at	3614.400	3563.220	2604.650	1945.710	3371.290	3406.660
117_at	1032.670	1164.150	510.692	5061.200	452.166	400.477
121_at	5917.800	6826.670	4562.440	5870.130	3869.480	3680.440
1255_g_at	224.525	395.025	207.087	164.835	111.609	130.123
1294_at	799.786	839.787	592.434	593.632	431.526	332.962
.....						
(22277 more rows)						

Nummerical summaries

In natural scale

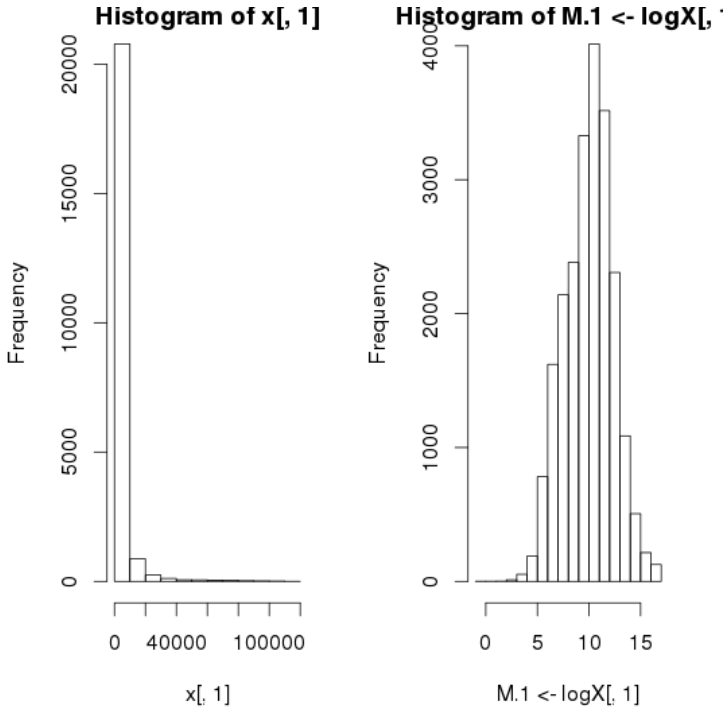
	DMSO013	Diox014	DMSO016	Diox018	DMSO020	Diox022
Min.	1	1	2	2	1	1
1st Qu.	327	317	192	233	172	169
Median	1146	1102	774	844	692	686
Mean	3459	3451	3364	3503	3438	3536
3rd Qu.	3052	3055	2968	2880	2853	2860
Max.	115400	112600	95910	115300	104000	119300

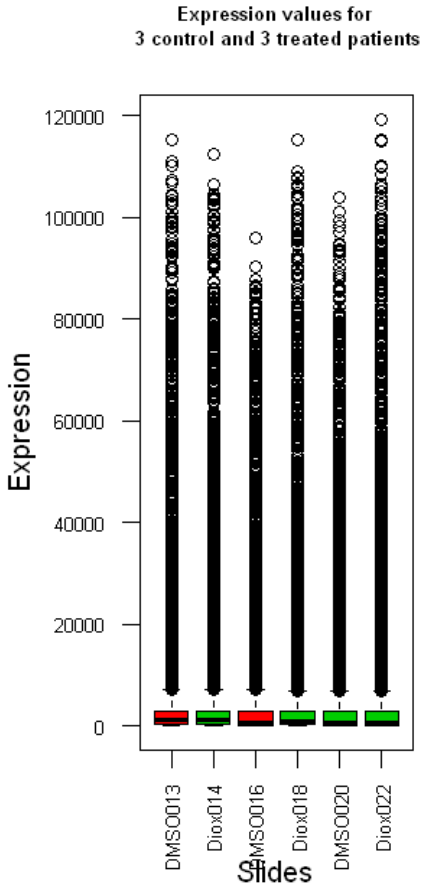
Data are clearly assymetric, better take logs

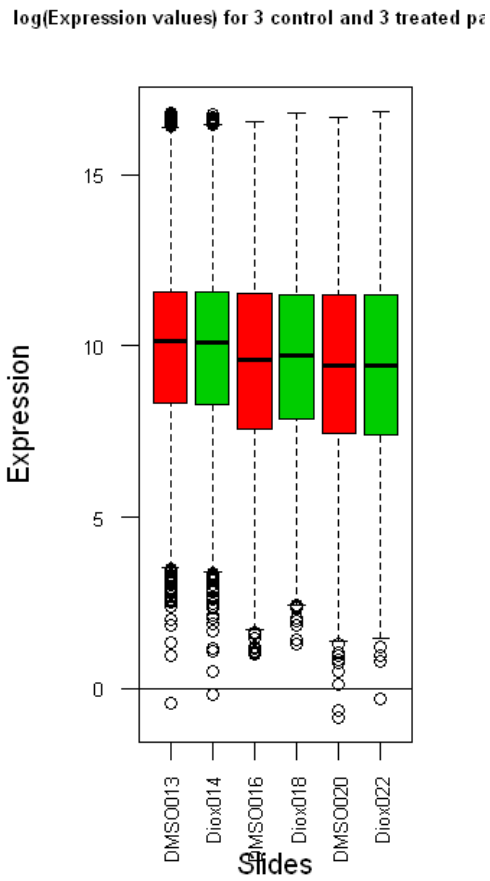
In log-scale

	DMSO013	Diox014	DMSO016	Diox018	DMSO020	Diox022
Min.	-0.443	-0.173	1.007	1.294	-0.871	-0.311
1st Qu.	8.352	8.310	7.584	7.863	7.429	7.405
Median	10.160	10.110	9.596	9.722	9.434	9.422
Mean	9.988	9.955	9.521	9.678	9.416	9.400
3rd Qu.	11.580	11.580	11.540	11.490	11.480	11.480
Max.	16.820	16.780	16.550	16.810	16.670	16.860

Plotting the data







From univariate to multivariate

- Univariate plots can give some insight
  - detect some atypical sample
  - show the need for transformations

  But relations between variables stay uncovered
- Multivariate analysis simultaneously looks at all the variables
- Doing this in reduced dimension allows
  - To rely on all variables
  - Keep only most informative dimensions

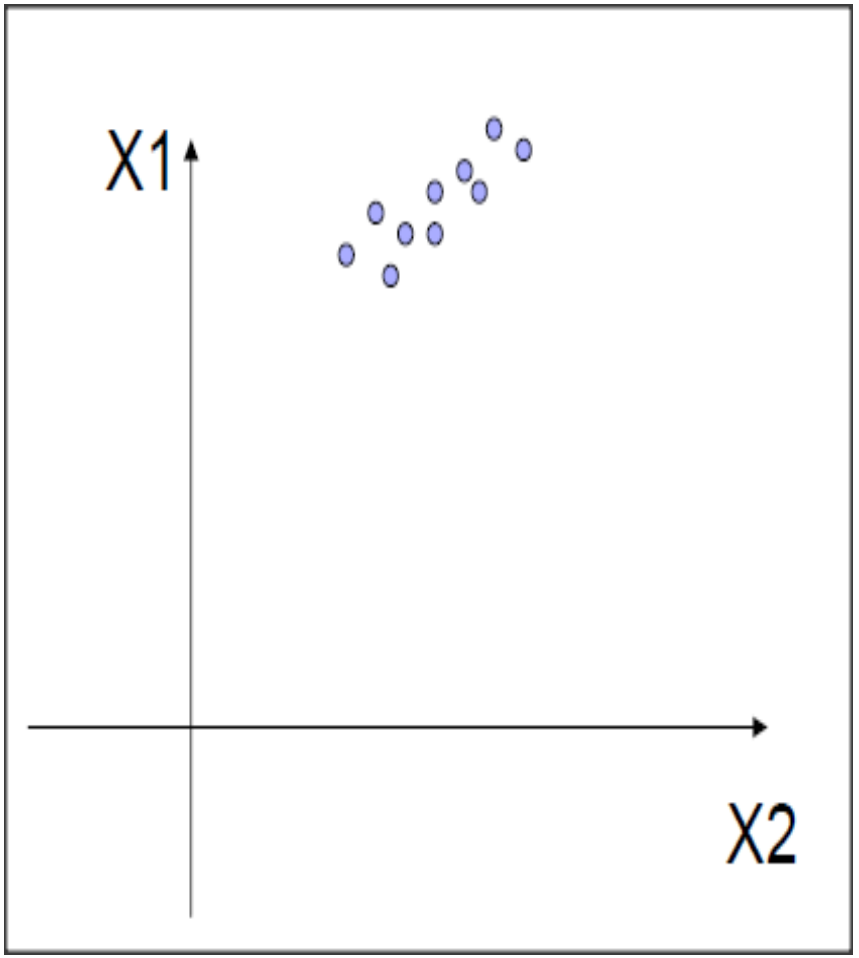
Principal Components Analysis

Principal Components Analysis

- Given a ***KxN*** data matrix containing ***K*** (correlated) measurements on ***N*** samples (objects/individuals…)
- Decomposes data matrix in new ***K*** components that
  - account for different sources of variability in the data,
  - are uncorrelated, that is each component accounts for a different source of variability,
  - have decreasing explanatory ability: each component explains more than the following
  - allow for a lower dimensional representation of the data in terms of scores on principal components.
  - get an overview of the dominant patterns and major trends in the data (visualize clusters, identify outliers)

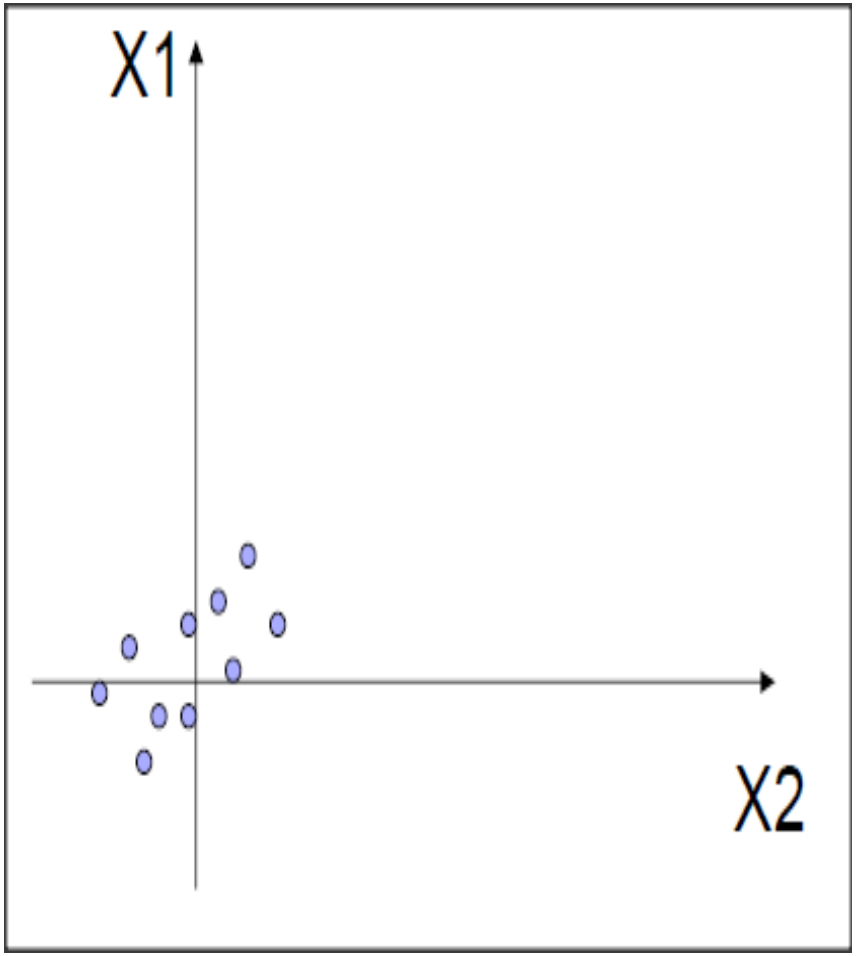
How does PCA work

- Assume a data matrix 2xN of two correlated variables.
- Being the data correlated it is difficult to separate each source of variability
- If K were much higher it would even be more difficult.



How does PCA work

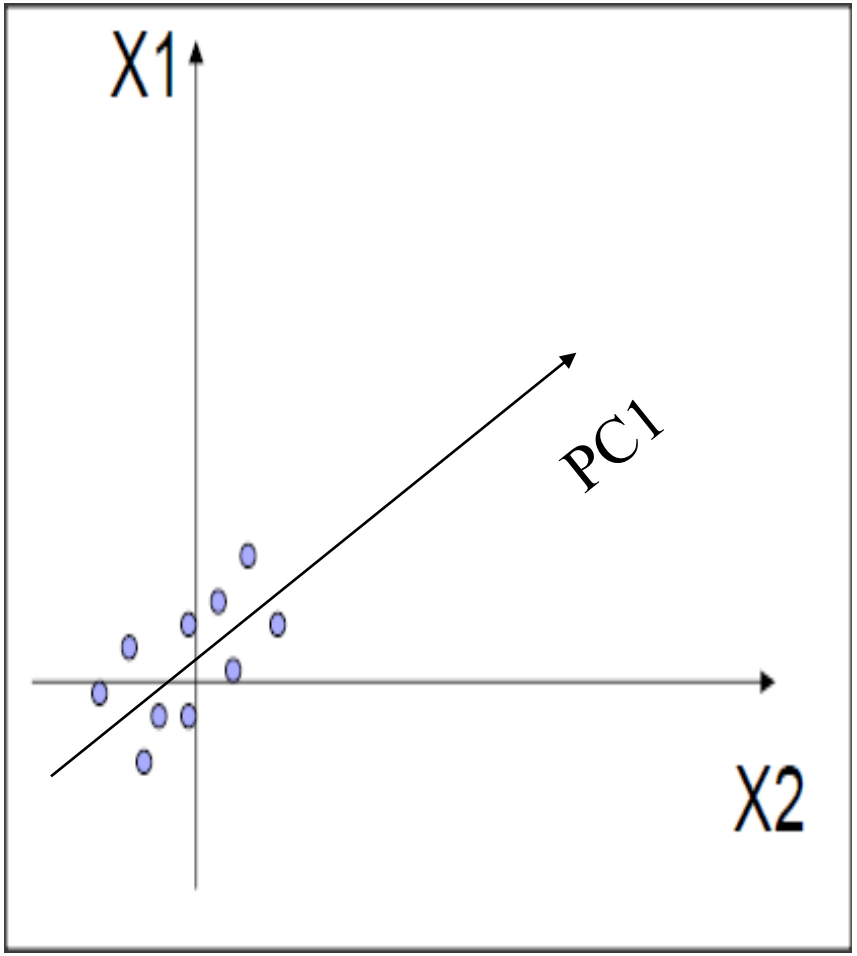
- Transform the data
  - Center each variable substracting its mean
  - Scale each variable dividing by its SD
- All variables are now comparable:
  - Mean = 0
  - SD = 1



How does PCA work

First principal component:

- a linear combination of
- all the original variables
- that goes along the direction of highest variability in the data
  - explains the *maximum amount of variation in the data*

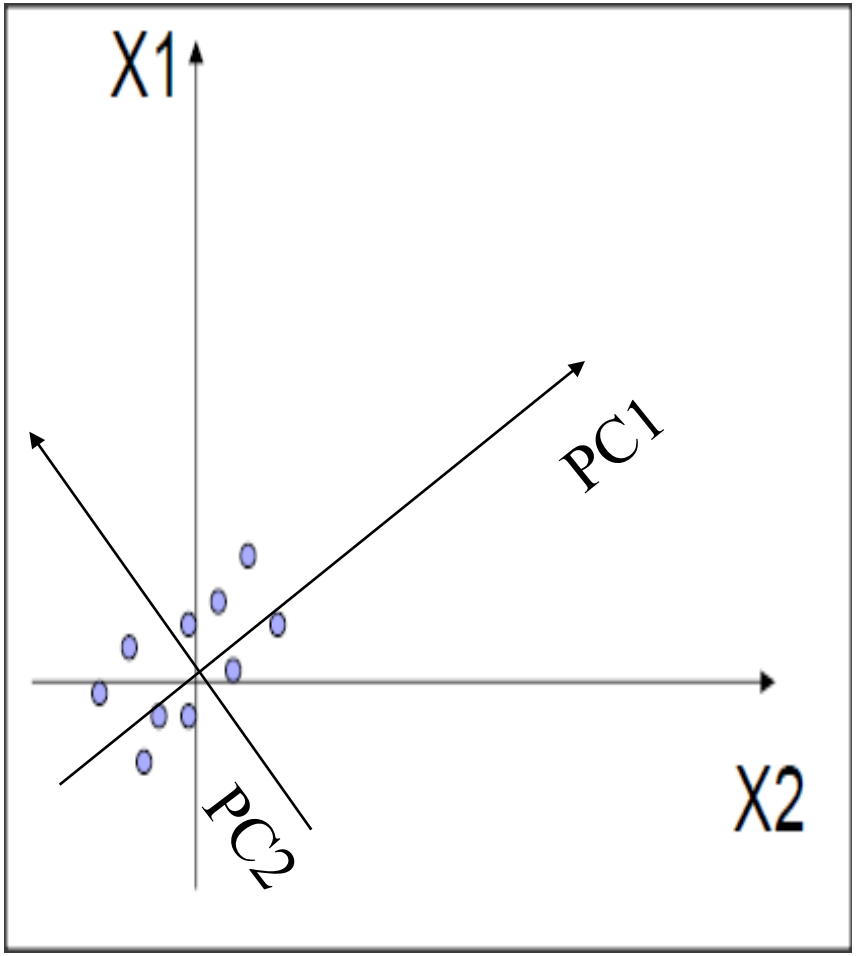


How does PCA work

2nd principal component:

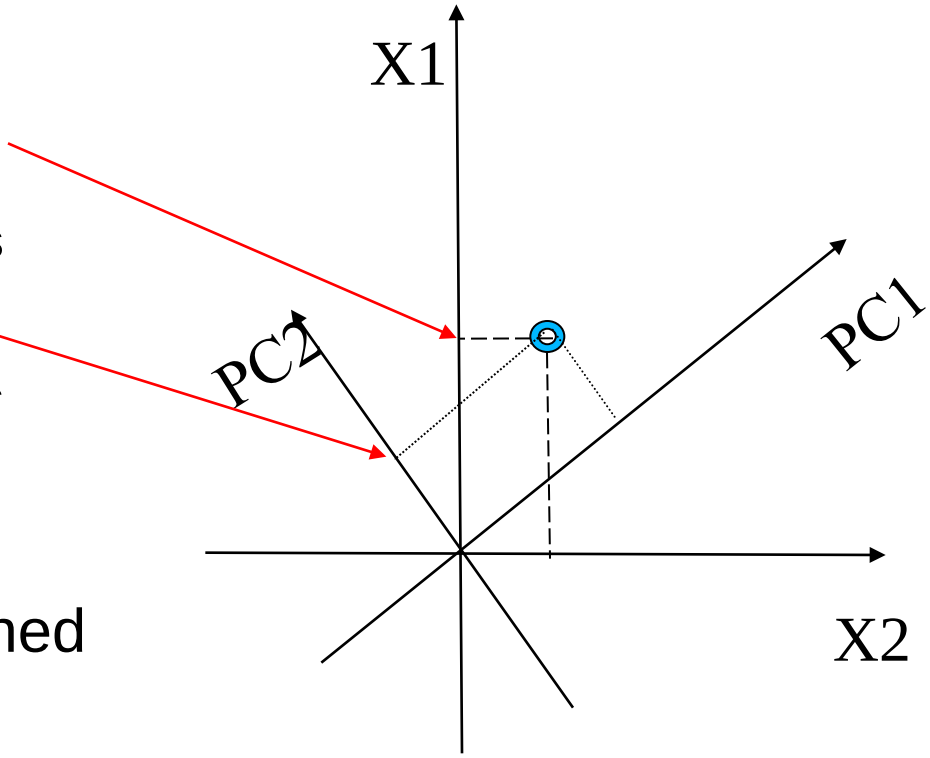
- a linear combination of
- all the original variables
- that goes along the next direction of highest variability in the data
  - ***orthogonally*** to first PC
  - explains the *maximum amount of* ***remaining*** *variation in the data*

- Successive PCs describe ***decreasing*** amount of ***remaining*** variation.



How does PCA work

- PCA provides a new set of coordinates for the observations
  - Original coordinates
    - •Value of the variables
  - New coordinates
    - •Value of PCs: ***scores***
- Scores are the new coordinates in the orthogonal system defined by PCs.

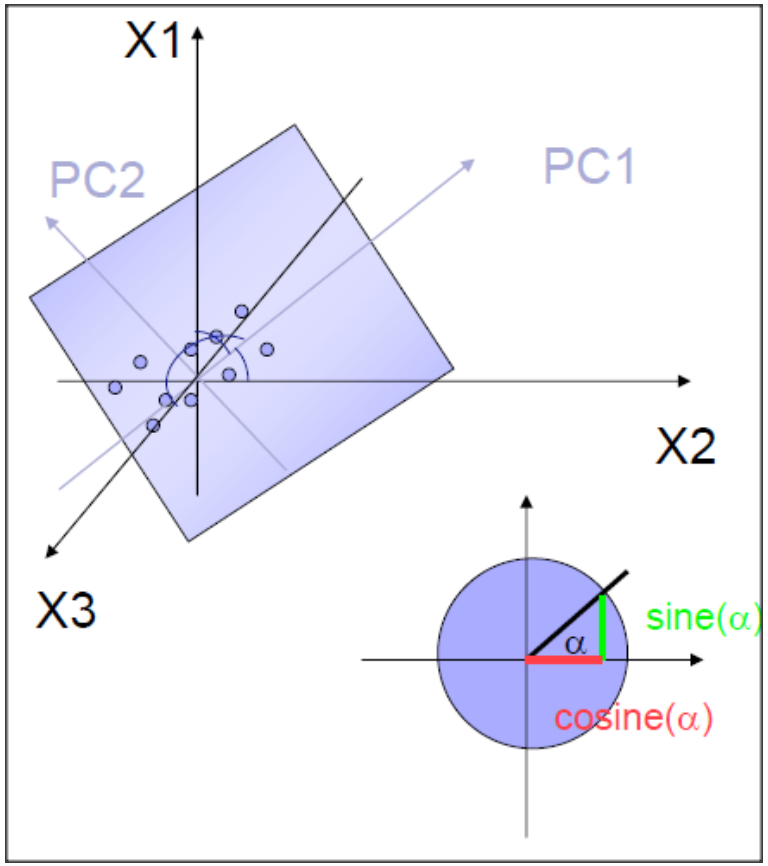


Representing data in the PCA space

- PCs have been derived so that
  - They are orthogonal
  - Each PC explains the *maximum amount of* ***remaining*** *variation in the data*
- This means that it is not necessary to use all PCs to visualize the data in this new coordinate system
  - Taking the first PCs will often explain a high percentage of variability.
  - Usually only first 2 or 3
  - ***This should always be checked!!!***

Interpreting PCs

- PCs can be interpreted by looking at which of the original variables contribute most to their variability
  - The more a variable is correlated with a PC the highest its influence.
- Size of contributions of each variable: *loadings*
  - Loadings are the *cosines of the angle between variables and PCs*



PCA - an example in R

```
iris.pc=prcomp(iris[,1:4], scale=T)

summary(iris.pc)
```

Importance of components:

	Comp.1	Comp.2	Comp.3	Comp.4
Standard deviation	1.708	0.956	0.383	0.144
Proportion of Variance	0.730	0.228	0.037	0.005
Cumulative Proportion	0.730	0.958	0.995	1.000

Loadings:

	Comp.1	Comp.2	Comp.3	Comp.4
Sepal.Length	0.521	-0.377	0.720	0.261
Sepal.Width	-0.269	-0.923	-0.244	-0.124
Petal.Length	0.580		-0.142	-0.801
Petal.Width	0.565		-0.634	0.524

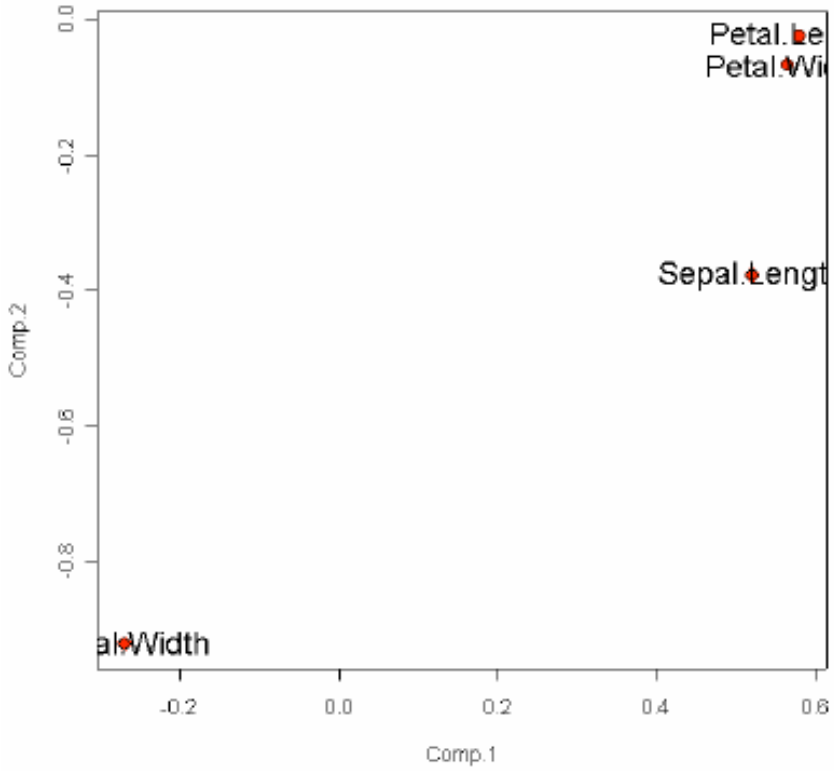
PCA - an example in R

```
names(iris.pc)
[1] "sdev"     "rotation" "center"   "scale"    "x"

iris.pc
Standard deviations:
[1] 1.7083611 0.9560494 0.3830886 0.1439265
Rotation:
                    PC1         PC2        PC3        PC4
Sepal.Length  0.5210659 -0.37741762  0.7195664  0.2612863
Sepal.Width  -0.2693474 -0.92329566 -0.2443818 -0.1235096
Petal.Length  0.5804131 -0.02449161 -0.1421264 -0.8014492
Petal.Width   0.5648565 -0.06694199 -0.6342727  0.5235971

iris.pc$sdev^2     #VARIANCE
[1] 2.91849782 0.91403047 0.14675688 0.02071484
iris.pc$sdev^2/sum(iris.pc$sdev^2) #proportion of var
[1] 0.729624454 0.228507618 0.036689219 0.005178709
```

PCA - an example in R

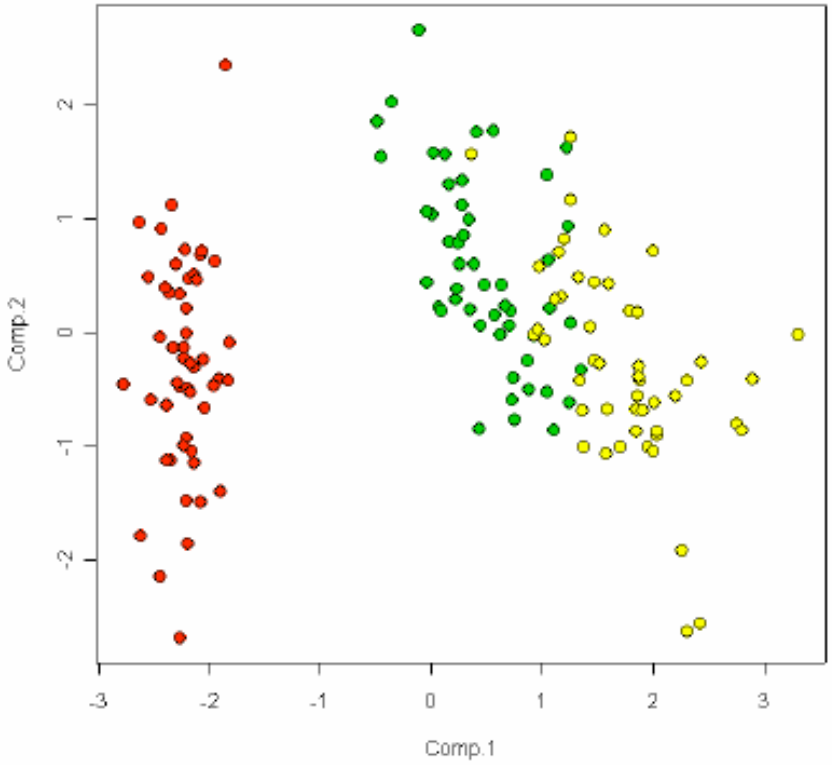
Loading plot

`plot(iris.pc$rotation[,1:2])`



Score plot

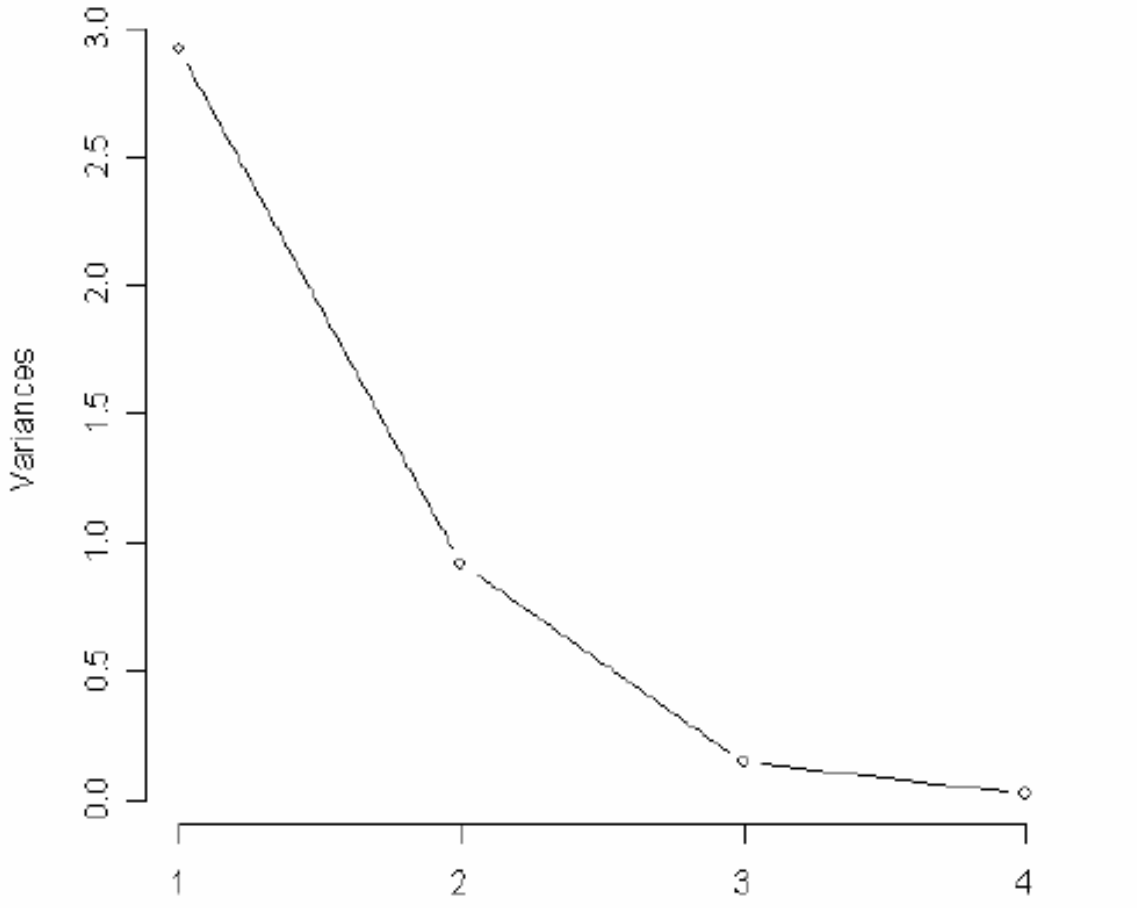
`plot(iris.pc$x[,1:2])`



PCA - an example in R

Scree plot

```
screeplot(iris.pc, type="line")
```



PCA – Algebraic basis

- PCA works by calculating a new system of coordinates.
- The directions of this new coordinate system are the eigenvectors of the covariance matrix A (p,p).
- An eigenvector of a matrix A is defined as a vector such as: $A\,\mathbf{z} = \lambda\,\mathbf{z}$

  where λ is a scalar called eigenvalue
- Each eigenvector has its own eigenvalue.
- An example:

PCA – Algebraic basis

- $A\,\mathbf{z} = \lambda\,\mathbf{z}$

$$A = \begin{bmatrix} -1 & 1 \\ 0 & 2 \end{bmatrix} \qquad \lambda_1 = -1 \qquad \lambda_2 = -2 \qquad z_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad z_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

$$Az_1 = \begin{bmatrix} -1 & 1 \\ 0 & 2 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 0 \end{bmatrix} = (-1) \cdot \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \lambda_1 z_1$$

$$Az_2 = \begin{bmatrix} -1 & 1 \\ 0 & 2 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ -1 \end{bmatrix} = (-2) \cdot \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \lambda_2 z_2$$

- In these cases, the new vector is just a (scalar) multiple λ of the original vector

PCA – Algebraic basis

- The values λ satisfying $Az = \lambda z$ are called eigenvalues of A.
- The corresponding nonzero vectors z are called eigenvectors.
- Consider the variance-covariance matrix A
- The eigenvalue with the largest absolute value (λ_1) will indicate that the data have the largest variance along its eigenvector, which is the first PC (PC1)
- The eigenvectors of A provide sets of coefficients defining p linear functions of the original variables that are PCs
- The PCs are orthogonal
- If A has eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$ then the PCs have variances $\lambda_1, \lambda_2, \ldots, \lambda_p$ and zero covariances

PCA – Algebraic basis

- In our example A **z1**:

```
cov(scale(iris[,1:4])) %*% iris.pc$rot[,1]
                [,1]
Sepal.Length  1.520730
Sepal.Width  -0.786090
Petal.Length  1.693934
Petal.Width   1.648533
```

- In our example λ1 **z1**

```
iris.pc$sdev[1]^2*iris.pc$rot[,1]
Sepal.Length  Sepal.Width Petal.Length  Petal.Width
    1.520730    -0.786090     1.693934     1.648533
```

PCA – Choosing the number of components

- There are a few ways to decide how many PCs to retain
- Some common methods are:
  - retain the number required to explain some percentage of the total variation (e.g. 90%)
  - number of eigenvalues > average (1 if correlation matrix is used)
  - look for 'elbow' in scree plot
  - compromise between these
- The scree plot shows proportion of variance (or just variance) explained by each component

PCA - step by step

- Problem formulation. Ensure that the data set contains enough relevant information to solve the problem.
- Plot the raw data.
- First runs. Center the data, try different pretreatment schemes. Spot the presence of outliers, determine the presence of clusters.
- Later runs.
- Inspection of PCs variance to determine the optimum number of PCs.
- Inspection of scores and loadings.

PCA - Cautions

- Sometimes used as a method for simplifying data because PCs associated with smaller eigenvalues have smaller variances and might therefore be 'ignored'.

- This assumption requires caution
- When variables are on different scales, it is customary to use the correlation matrix (rather than the covariance matrix).
- These two formulations give different results : the eigenvalues for the two matrices are not related in a simple way. Theory not simple for correlation-based PCA.

PCA -Summary

T score matrix X data matrix P loading matrix

E residual matrix

- $X(n,p) = T(n,M)\, P'(M,p) + E(n,p)$

- Scores: are combination of the original variables and describe decreasing amount of variation
- Loadings: inform us of the magnitude and the manner of the contribution of each original variables to the scores

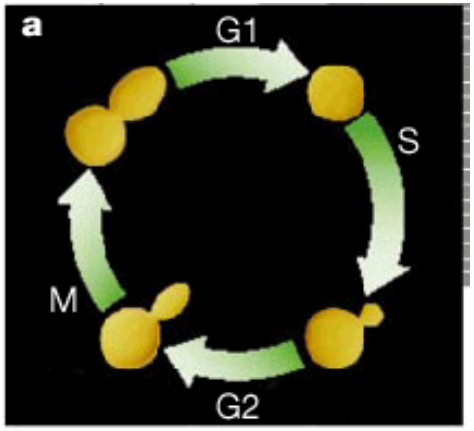
PCA in summary (1)

- PCA performs a transformation into a new set of orthogonal coordinates with decreasing ability (most, 1st PC, to least, last PC) to explain the observed variability.
- PCA analysis provides
  - % of variance explained by each PC
  - Loadings: Correlations between PCs and variables
    - Use these to (try to) interpret what the PCs mean
  - Scores: Values of the observations in the PC system of coordinates
    - Use these to plot the observations in reduced dimension.

PCA in summary (2)

- PCA can be seen as a method to unravel underlying trends or patterns in the data
  - The concept of *latent variable* is related with it
- PCA can be used to provide an overview of the data to reveal
  - Dominating variables
  - Trends
  - Patterns such as outliers, groups, clusters
  - Similarities/Dissimilarities

Transcriptional regulation and function during the human cell cycle, Cho et al. (2001) *Nature Genetics* Vol 27, 48-54

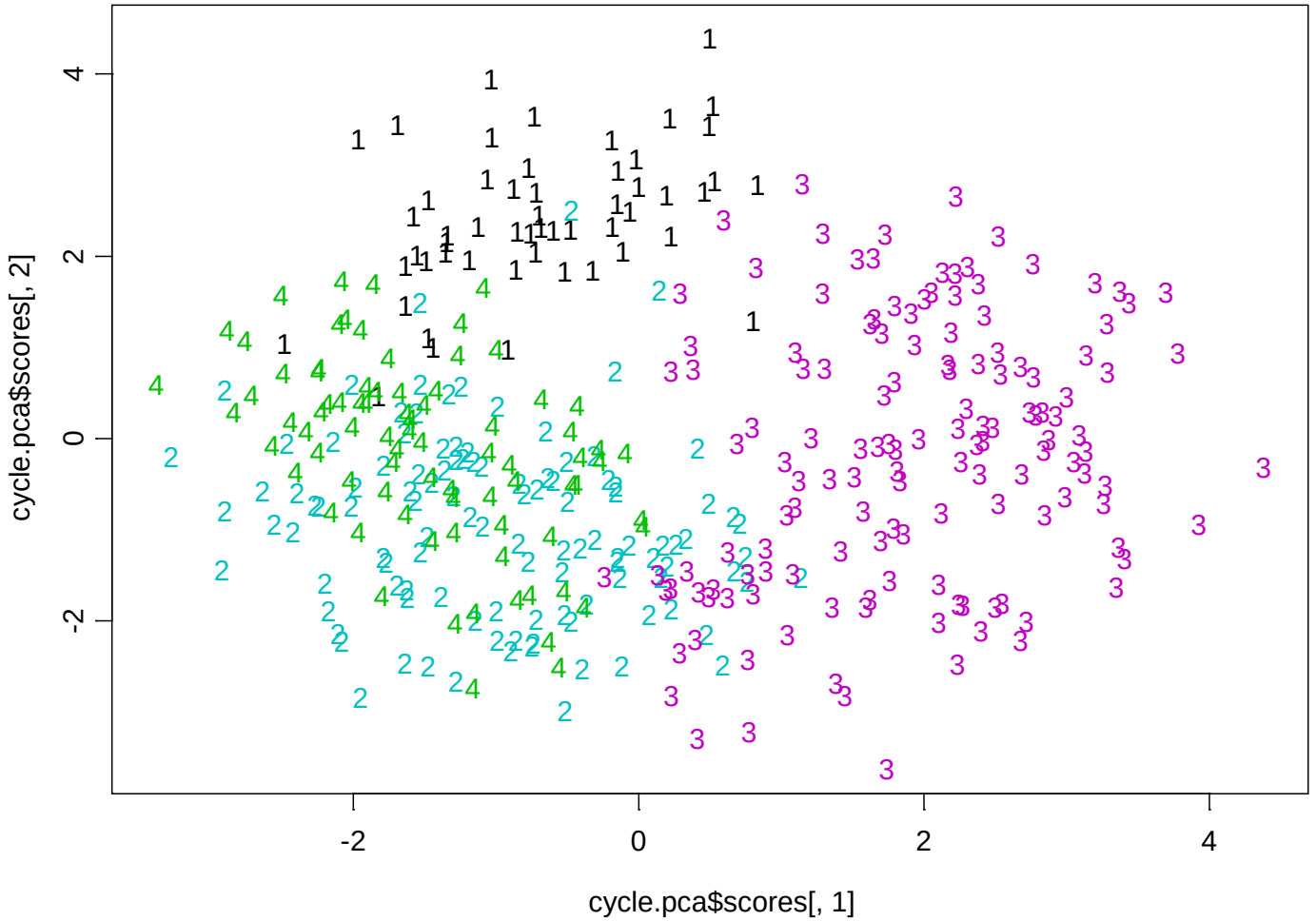


- Goal of the study was *to identify cell-cycle–regulated transcripts in human cells*
- Primary fibroblasts prepared from **human foreskin** were grown to approximately 30% confluence and **synchronized in late G1** using a double thymidine-block protocol9. Cultures were then released from arrest, and cells were collected every 2 hours for 24 hours, covering nearly **2 complete cell cycles**.
- Messenger RNA was isolated, labeled and hybridized to sets of arrays containing probes for approximately 40,000 human genes and non-overlapping ESTs. We carried out the entire synchronization experiment in duplicate under identical conditions for 6,800 genes on Affy array. The two data sets were averaged and analyzed using both supervised and unsupervised clustering of expression patterns.
- **Most genes were upregulated in one of the phases of the cell cycle so they were labelled accordingly to this as *G1→1; S→4; G2→2; M→3;***

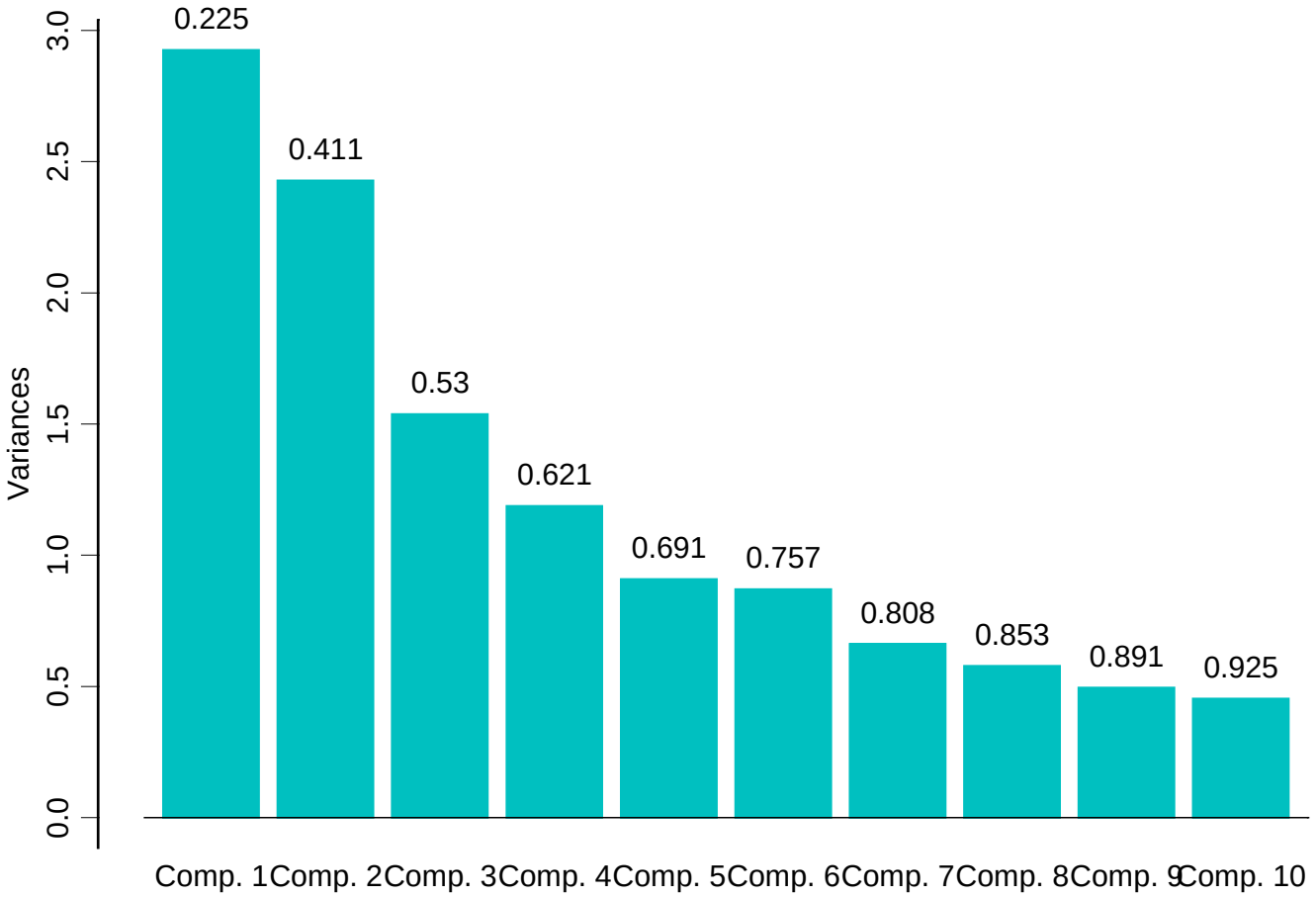
PCA analysis of cell cycle stages

PCA projection: 387 genes in 13-dim space (time points) are projected into 2D space using correlation matrix.

Gene phase: ***G1→1; S→4; G2→2; M→3;***



Variance in data explained by the first *n* principal components



PCA projection: 13 samples (time points) in 387-dim space (genes) are projected to 2D space using correlation matrix;

Each sample is labeled by its time point