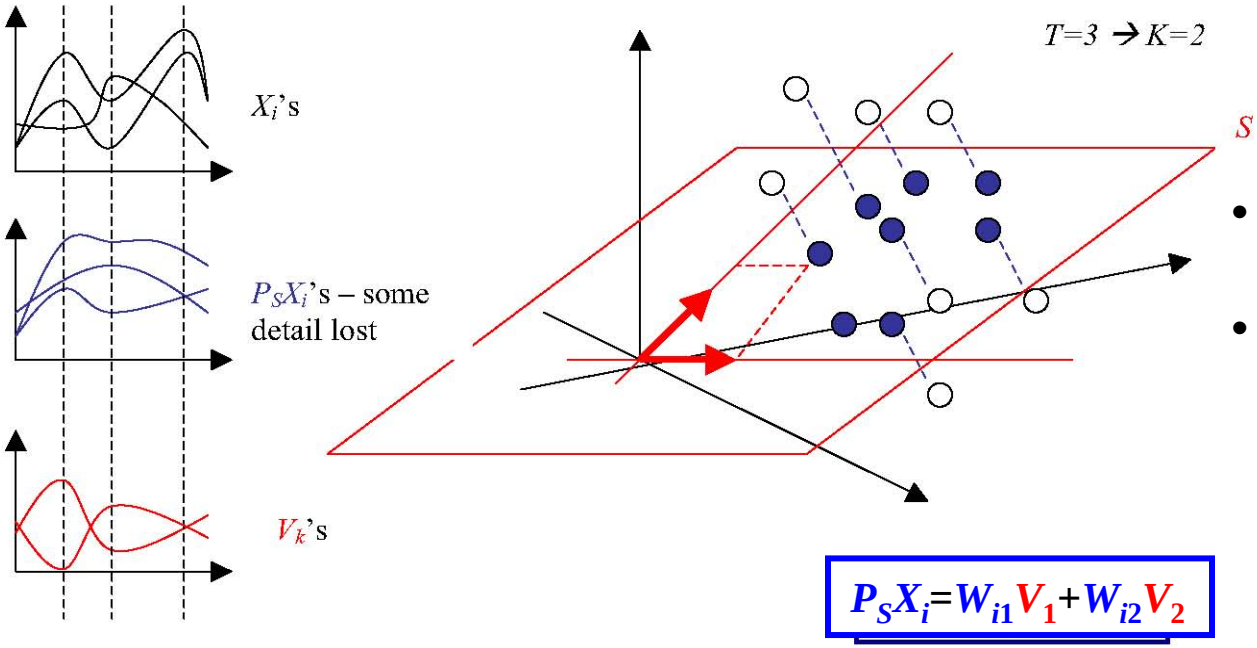
Basic Patterns in Data and (Unsupervised) Dimension Reduction: Principal Components Analysis.

Principal Components (PCA or equivalently Singular Value Decomposition)

p points in R^T (p = # of genes, T = # of conditions)

- Extract a few basic expression patterns (find a subspace).
- Give a low-dimensional reconstruction of the gene expression profiles (project the points)



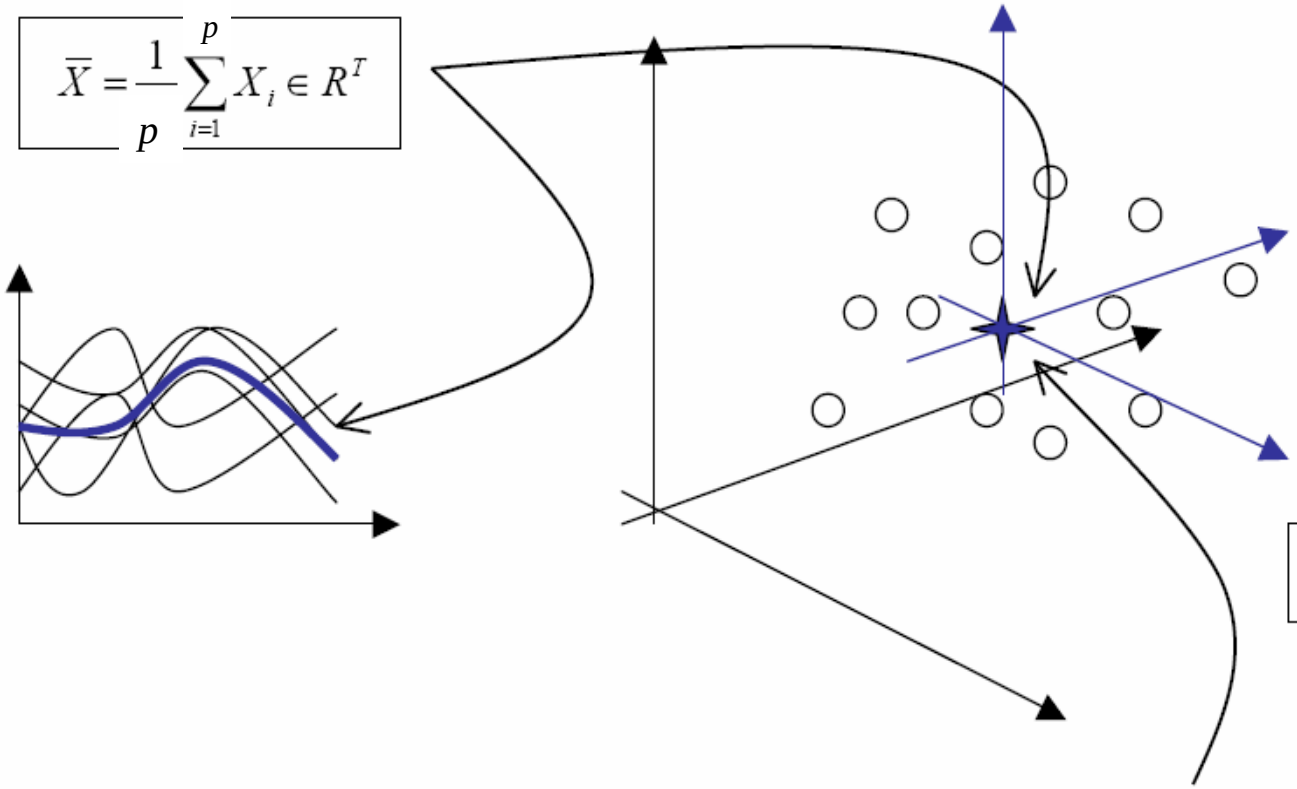
- As a "structural summary" of the data
- As a "cleaning" step prior to further analyses

$$P_S X_i = W_{i1} V_1 + W_{i2} V_2$$

"Structural" feature of interest i.e. criterion: Variability of the data cloud
(look for a subspace that captures a large share of it)

For the purpouse of investigating the variability structure, is does not matter where the data cloud is centered (mean vector; average expression profile)

$$\bar{X} = \frac{1}{p}\sum_{i=1}^{p} X_i \in R^T$$



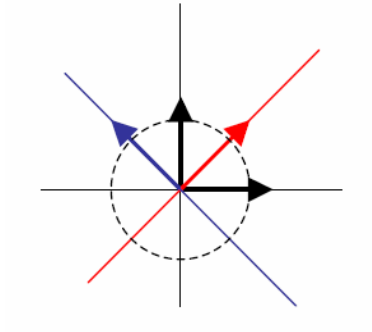
translate the origin

Thus performing the analysis on the X's, or the X's centered by column is the same

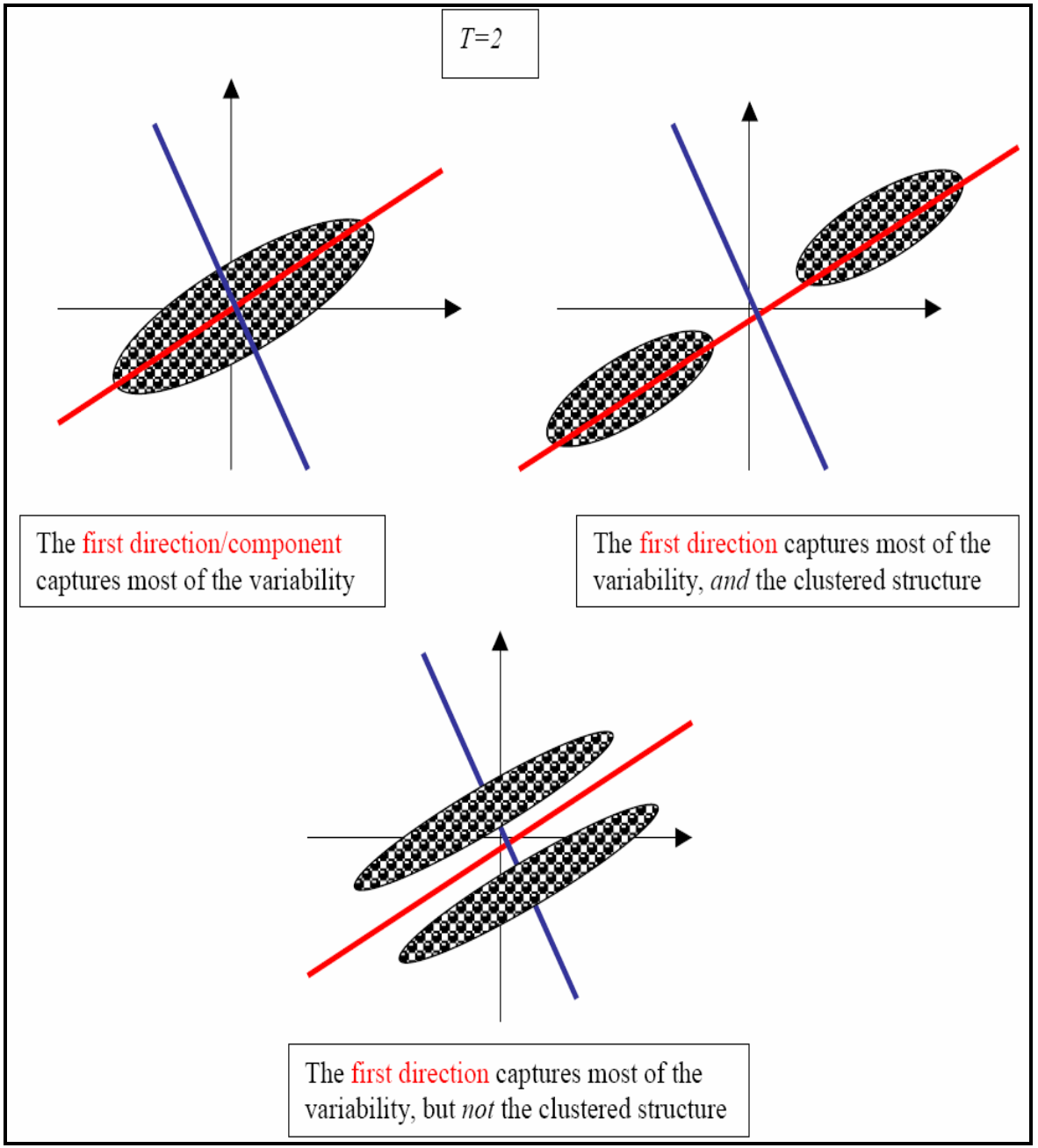
$X_i, i = 1..\ p$

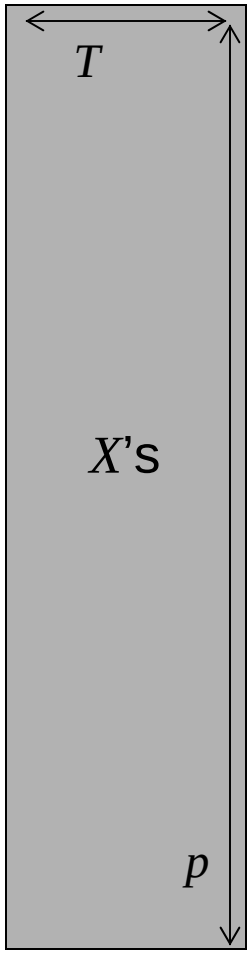
$(X_i - \bar{X}), i = 1..\ p$

First: determine a set of T orthogonal directions, ordered in terms of variability displayed by the data along them. "Natural axes" of the data cloud, ordered in terms of spread. A rotation of the original variables/axes

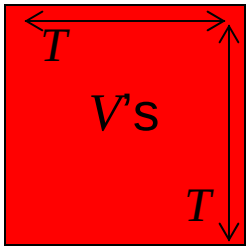


PCA is an "exhaustive" dimension reduction tool for Gaussian data: all there is to structure is center ("translated out") and variability – no odd shapes, no clusters, no holes. Not so for data whose structure is more complicated, but PCA can still be applied as a tool aiming at variability alone.





Eigenvct's of Σ_X



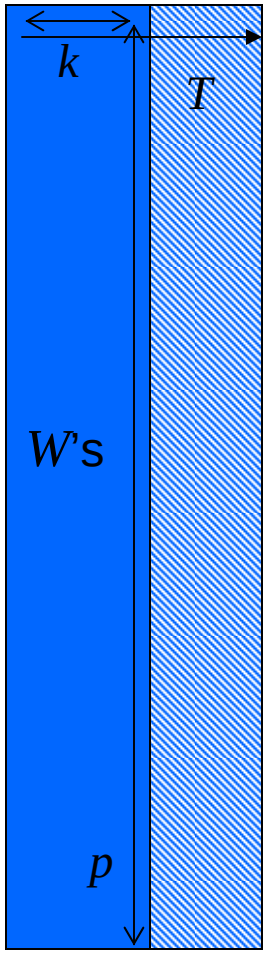
Rows: directions in R^T ordered by variability

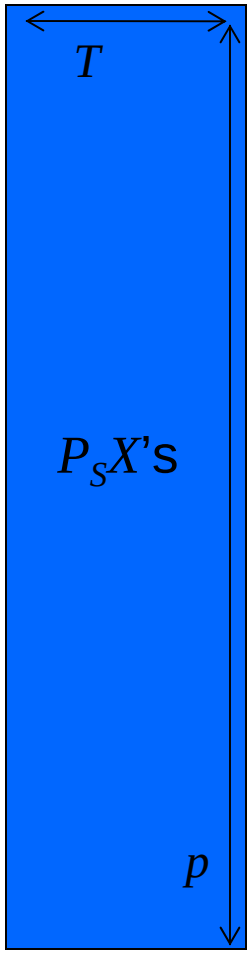
Select k

Low-dimensional representation

Clmns: lin comb's of X's (clmns) ordered by variab

Rows: lin comb's of selected V's (rows)





Spectral decompositions of the covariance matrix

$$\Sigma_X = \frac{1}{p}\sum_{i=1}^{p}(X_i - \overline{X})(X_i - \overline{X})' = \sum_{j=1}^{T}\lambda_j v_j v_j'$$

$\{v_1 \dots v_T\}$ Eigenvectors, orthogonal directions ranked in terms of…

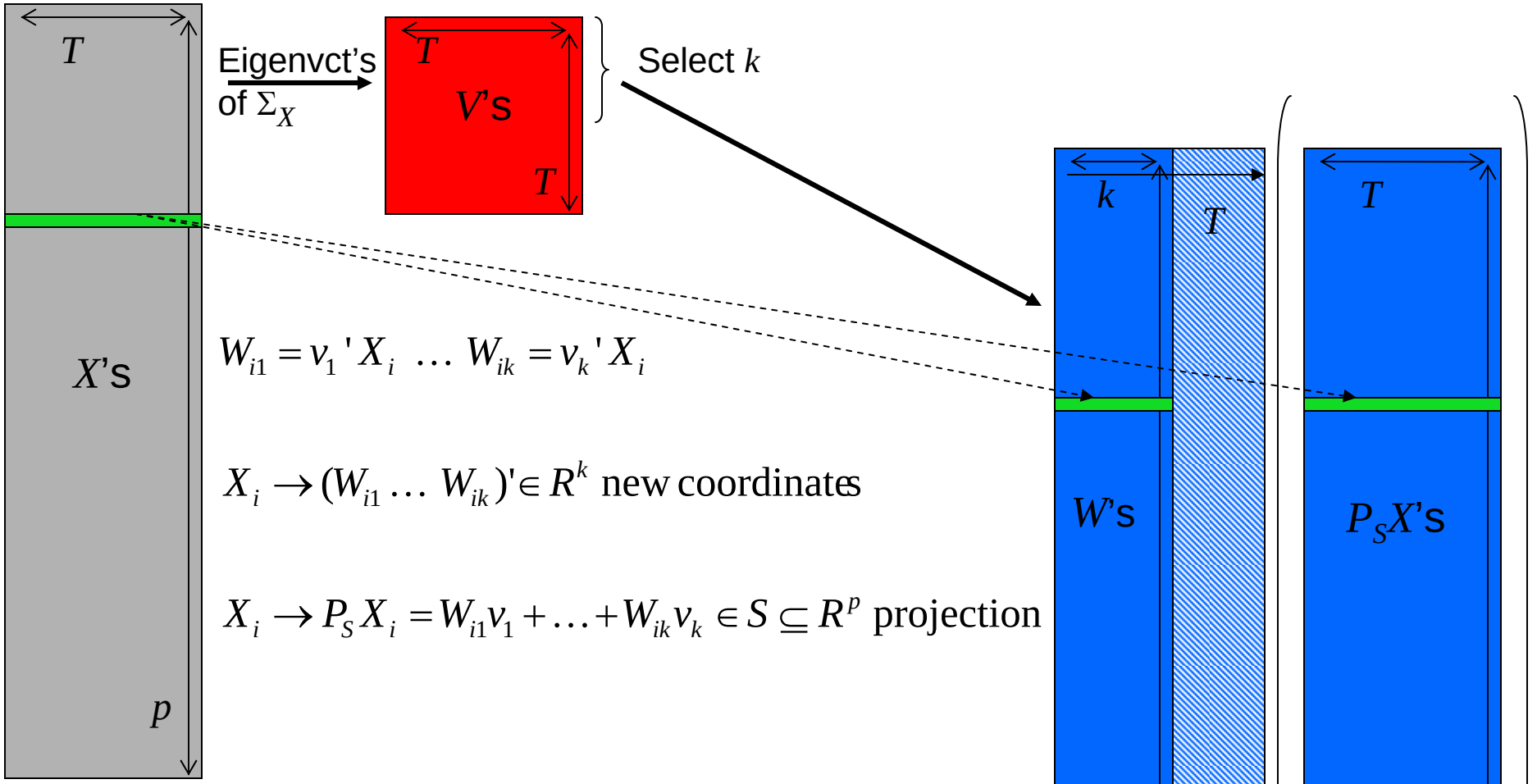
$\lambda_1 \geq \lambda_2 \dots \geq \lambda_T$ Eigenvalues, variability (real ≥ 0)

$S = Span(v_1 \dots v_k)$

$W_{i1} = v_1' X_i \quad \dots \quad W_{ik} = v_k' X_i$

$P_S X_i = W_{i1} v_1 + \dots + W_{ik} v_k$

k-dimensional projection preserving most of the variability structure of the data cloud.



$$W_{i1} = v_1' X_i \ \ldots \ W_{ik} = v_k' X_i$$

$$X_i \to (W_{i1} \ldots W_{ik})' \in R^k \text{ new coordinates}$$

$$X_i \to P_S X_i = W_{i1} v_1 + \ldots + W_{ik} v_k \in S \subseteq R^p \text{ projection}$$

A real simplification/reduction occurs if we actually have that, as a group

$$X_i \approx \sum_{j=1}^{K} W_{i,j} V_j = P_{Span(V_1 \ldots V_K)} X_i \quad , \quad K < T$$

Second: determine how many directions to retain (k).

1. Consider the proportion of explained variability, and retain as many directions as needed to explain a selected proportion

$$\frac{\lambda_1}{\sum_{j=1}^{T}\lambda_j} \quad \frac{\lambda_2}{\sum_{j=1}^{T}\lambda_j} \quad \cdots \quad \frac{\lambda_T}{\sum_{j=1}^{T}\lambda_j}$$

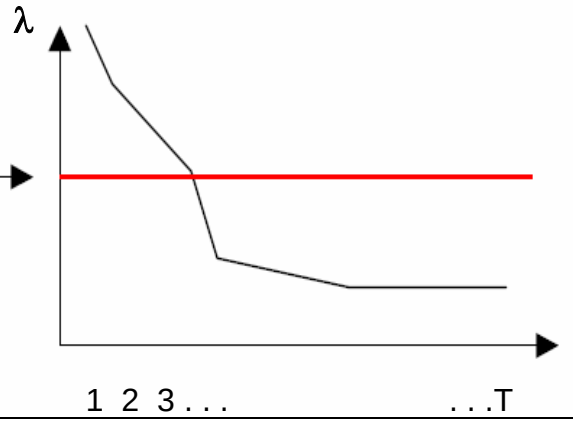
← One by one

$$\frac{\lambda_1}{\sum_{j=1}^{T}\lambda_j} \quad \frac{\lambda_1+\lambda_2}{\sum_{j=1}^{T}\lambda_j} \quad \cdots \quad \frac{\sum_{j=1}^{T}\lambda_j}{\sum_{j=1}^{T}\lambda_j} = 1$$

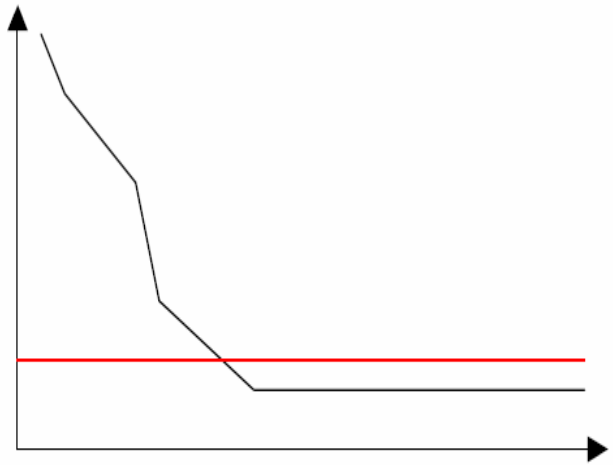
← Cumulative. Stop when you reach, say, .80 i.e. 80%

2. Consider the average explained variability per component, and retain directions with an explanatory capability above average – on the scree plot:

$$\overline{\lambda} = \frac{1}{T}\sum_{j=1}^{T}\lambda_j$$



3. Look for bends in the scree plot. If there is a clear bend, keep directions associated with eigenvalues before the bend – those afterwards have comparable, small(er) size (smaller the more they are)



4. "Testing" version of 3. *If the data is elliptical (Gaussian or about)*, we can perform a sequence of tests to assess how many tail eigenvalues are statistically equal to one another. This is based on a large N chi-square null distribution (not clear whether normality is a viable assumption)

$$H_0: \quad \lambda_{T-h+1} = \ldots = \lambda_T$$

$$H_1: \quad not$$

$$\left(N - \frac{2T-11}{6}\right)\left(h \ln \bar{\lambda} - \sum_{j=T-h+1}^{T} \ln \lambda_j\right) \overset{approx}{\sim} \chi^2_{\frac{1}{2}(h-1)(h+2)}$$

Note: Many, and more sophisticated approaches exist; large stat literature

Think about:

- **"scrambling" scheme to provide a chance background for the choice of K ("reference curves" on a scree plot).**
- **Re-sampling scheme to assess the stability (sampling variability) of eigenvectors (main variability directions) and eigenvalues (shares of explained variability).**

What are we missing?

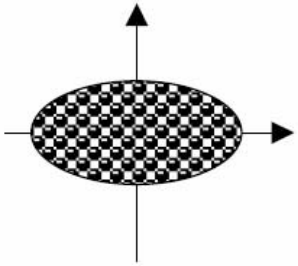
- With 1, we fix the percentage of variability we are missing
- With 2,3 and 4 we don't. It might not be negligible
- With 1 and 2, there might still be variability structure in the neglected directions
- With 3 and 4, the variability is approximately the same in all neglected directions

If the percentage of variability we are missing is very small (e.g. 2%), we might argue that whatever structural feature might live in the neglected directions, it occurs on a scale so small that we do not care.

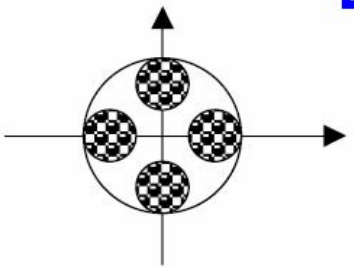
But if the percentage of variability we are missing is not very small (e.g. 20 or 30%), we ought to investigate what is going on there.

When it looks "structured", is what we are throwing away "non-experimental" structure? Can we call it such if it plays out on small scale? Can we identify it as such even if it plays out on sizeable scale?

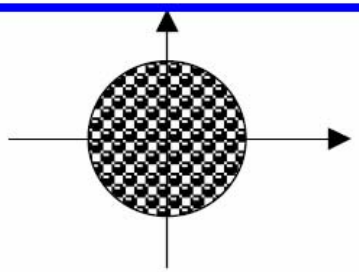
For example, if neglecting last two directions:



Still some variability structure



No variability structure… but some structure!



Spherical scatter, "noise-like": is this what we mean by no structure?

PCA/SVD analyses applied to microarray data sets:

• Holter N.S., Mitra M., Maritan A., Cieplak M., Banavar J.R., Fedoroff N.V. (2000). Fundamental patterns underlying gene expression profiles: Simplicity from complexity. PNAS 97: 8409-8414.

• Alter O., Brown P.O., Botstein D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. PNAS 97: 10101-10106.

• Alter O., Brown P.O., Botstein D. (2003). Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. PNAS 100:3351-3356.

Since the following example is a time course, also relevant analyses designed for time series or data with an intrinsic ordering structure:

• Peddada S. D., Lobenhofer L., Li. L., Afshari C., Weinberg C., Umbach D. (2003). Gene selection and clustering for time-course and dose-response microarray experiments using order-restricted inference. Bioinformatics 19:834-841.

• Liu D., Umbach D., Peddada S. D., Li L., Crockett P., and Weinberg C. (2004). A random-periods model for expression of cell-cycle genes. PNAS 101: 7240-7245.

An Example:

Yeast cell-cycle data

• Spellman P.T., Sherlock G., Zhang M.Q., Vishwanath R.I., Anders K., Eisen M.B., Brown P.O., Botstein D. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces Cerevisiae by microarray hybridization. Molecular Biology of the Cell 9: 3273-3297.

Preprocessed data set:

• j=1…T=12 (first of 15) times covering 2+ cycles; cdc (synch technique) course

• spotted arrays (thousands of yeast genes)

• normalized (to green) log-ratios

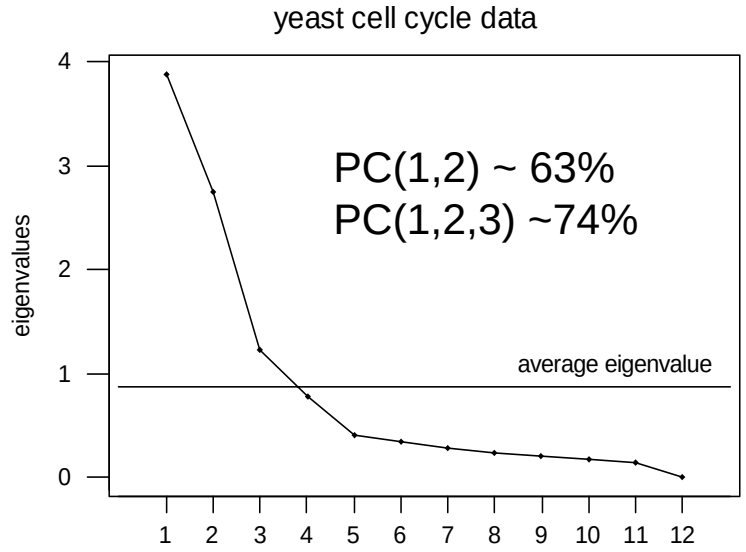
$$X_{ij} = \log\left(\frac{R_{ij}}{\rho_j G_{ij}}\right) \qquad \rho_j = \frac{\sum_i R_{ij}}{\sum_i G_{ij}}$$

• i=1…p=679 genes, selected for exhibiting periodic behavior (Fourier analysis) and having no missing entries

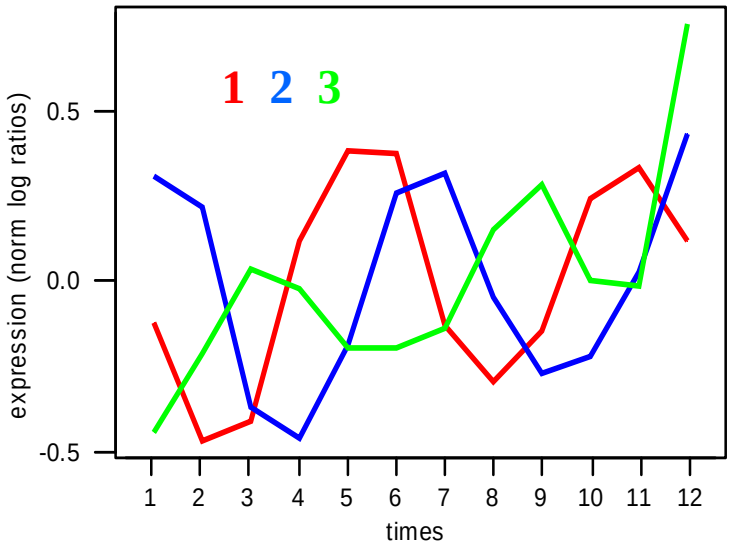
• gene (row) centering and standardization: eliminate average expression and variation magnitude effects; restrict analysis to “pure shapes” of gene expression profiles

$$X_{ij} \leftarrow \frac{X_{ij} - \overline{X}_i}{sd_i}$$

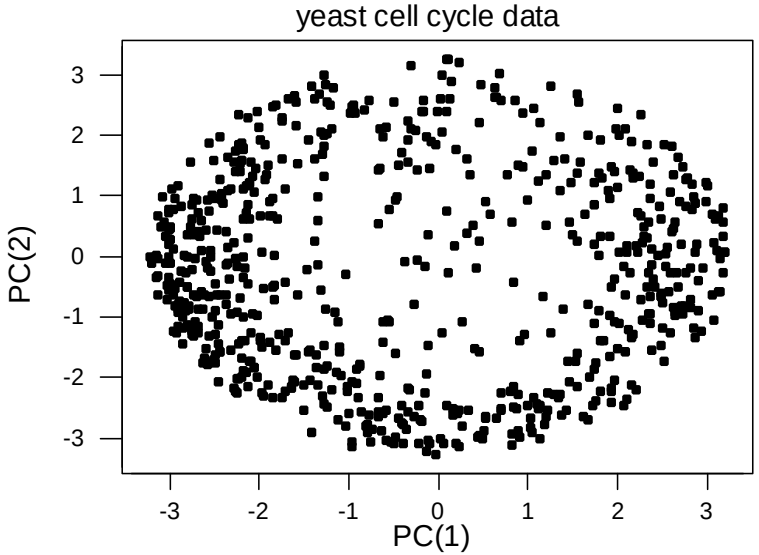
How complex is the data? (dimension)



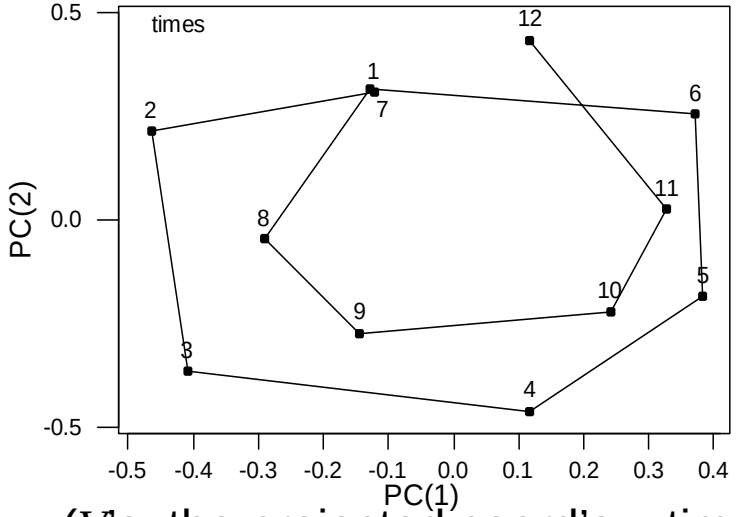
Basic patterns (eigenvectors)



Low dimensional representations (visualization)

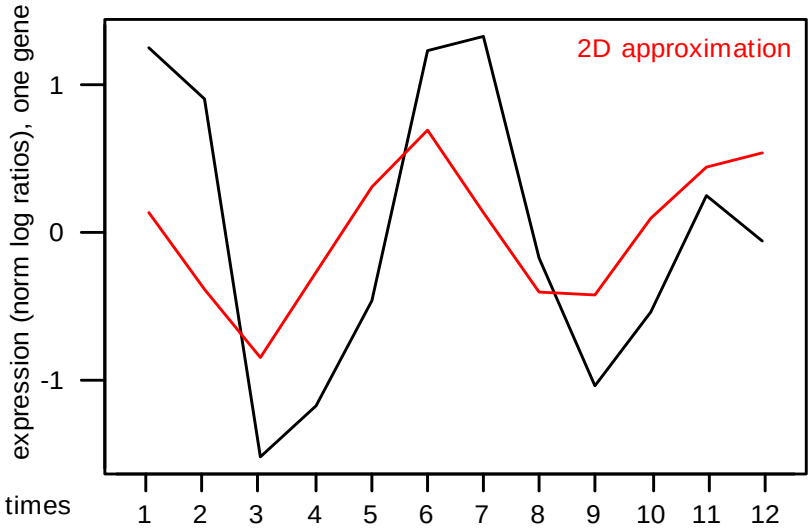


(*W*'s; the projected points – genes)

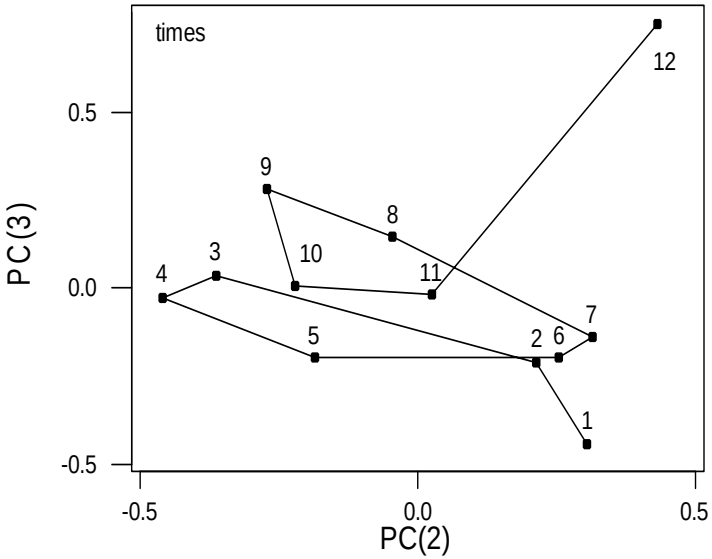


(*V*'s; the projected coord's – times)

Principal components for reducing noise and artifacts



($P_S X$'s, one instance)



Eliminate third+ PC directions to emiliorate dampening in amplitude and trend, in the assumption that they were artifacts:

- dis-synchronization
- expression reaction to synchronization drugs, “crowd” effects?

Using principal components to rank (and select) genes:

$$X_i = \sum_{j=1}^{T} W_{i,j} V_j$$

How close is X_i to V_j ?

$$P_{V_j} X_i = (W_{i,j}) V_j$$

$$\frac{\| P_{V_j} X_i \|^2}{\| X_i \|^2} = \frac{(W_{i,j})^2}{\sum_{j=1}^{T} (W_{i,j})^2} = corr^2(X_i, V_j) = R^2_{X_i | V_j}$$

(recall one norm is =1)

How close is X_i to $V_1 \ldots V_K$ as a group, i.e. to $Span(V_1 \ldots V_K)$?

$$P_{Span} X_i = \sum_{j=1}^{K} P_{V_j} X_i = \sum_{j=1}^{K} (W_{i,j}) V_j$$

$$\frac{\| P_{Span} X_i \|^2}{\| X_i \|^2} = \frac{\sum_{j=1}^{K} \| P_{V_j} X_i \|^2}{\| X_i \|^2} = \frac{\sum_{j=1}^{K} (W_{i,j})^2}{\sum_{j=1}^{T} (W_{i,j})^2} = R^2_{X_i | V_j \ldots V_K}$$

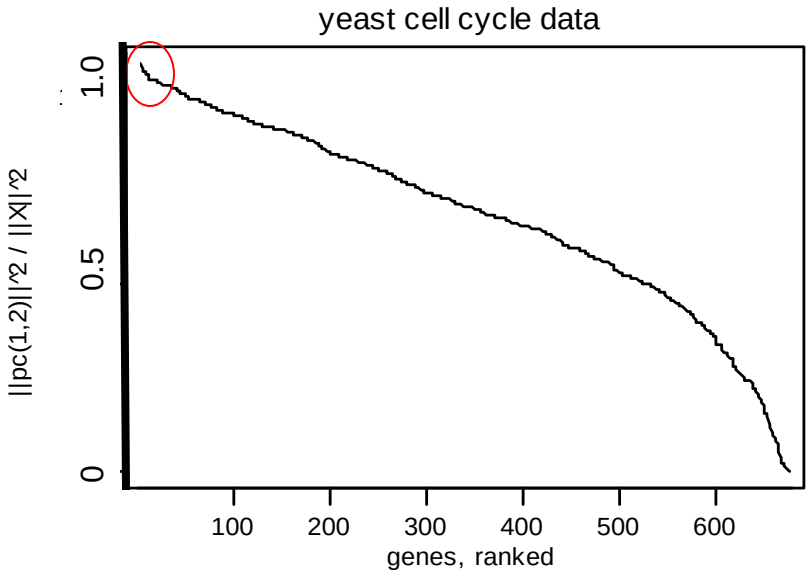
Squared correlation between ith gene (actual row, profile) and jth PCA (synthetic row, basic pattern)

Determination coefficient from a linear ols fit (ith gene on one or K PCA's, using the T conditions)

Genes closest to the natural direction(s) of highest variability in expression
Gene profiles better reconstructed in terms of basic expression patterns:

Identifying genes that "drive" patterns (ranking on projections)

(*W*'s; relative "size" of projections)



Genes closest to "pure" cycling behavior

YLR190W **top ORFs**
YOR391C
YKR037C
YML058W
YHR005C
YDR191W
YKL185W
YNL058C
YGR042W
YLR326W …

Think about
- "scrambling scheme" to provide a chance background ("reference")
- re-sampling scheme to assess the stability (sampling variability) for this ranking plot.

Using the correlation matrix: Perform the spectral decomposition of

$$R_X = \begin{pmatrix} 1 & cor(X_1, X_2) & \dots & cor(X_1, X_T) \\ cor(X_2, X_1) & 1 & \dots & cor(X_2, X_T) \\ \vdots & & & \vdots \\ cor(X_T, X_1) & cor(X_T, X_2) & \dots & 1 \end{pmatrix} = Diag(1/\sigma_j)\, \Sigma_X\, Diag(1/\sigma_j)$$

Remarks:

Same as var/cov matrix of the data after standardizing by column. Free of scale and measurement units.

Eigenstructures of R and Σ are different (eigenvalues and vectors). Average eigenvalue of a correlation matrix is always 1.

PCA applied to R captures only linear interdependencies among variables, not their spread (all standardized to sd = 1).

Do different variability scales matter for the analysis at hand? If not, use R.

For microarray data there isn't an issue of units of measure (all clmns have the same).

Principal components in R:

```
> library(stats)
> # two functions to perform PCA, prcomp() and princomp()
> help(prcomp)
```

prcomp(x, retx = TRUE, center = TRUE, scale. = FALSE, tol = NULL)

Arguments:

x: matrix (or data frame) with the data
retx: logical, whether rotated variables should be returned.
center: logical, whether variables should be shifted to zero center.
scale.: logical, whether variables should be scaled to unit variance. Default is FALSE; var/cov matrix. TRUE; correlation
tol: value below which components should be omitted (omitted if st. dev. ≤ 'tol' times st.dev. Of first component.)

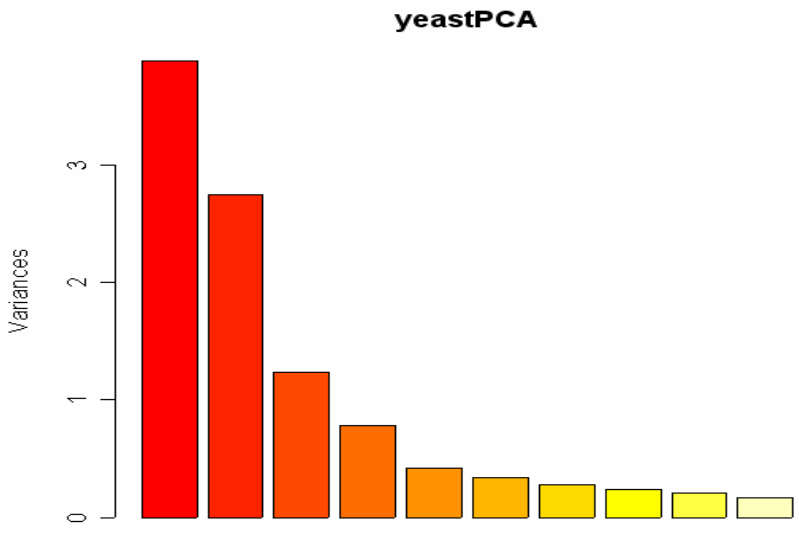
Calculation is done by singular value decomposition (more numerically stable, and mathematically equivalent.)

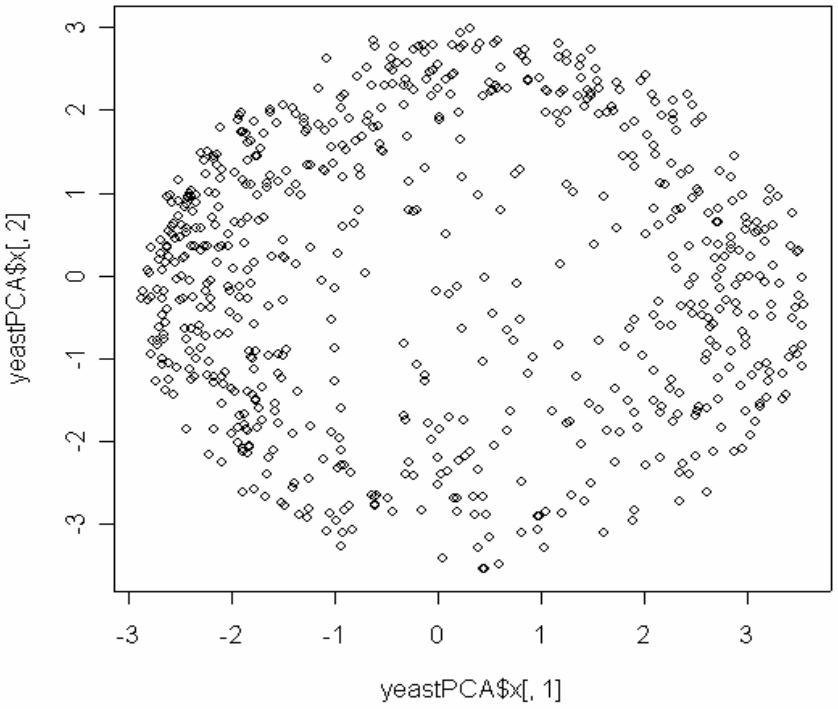
Value, list containing the following:

sdev: st. devs of princ comps (i.e. sq roots of eigenvalues). (λ's)
rotation: matrix whose columns contain the eigenvectors. (**V**'s)
x: if 'retx' is TRUE, matrix of rotated data (i.e. data multiplied by 'rotation' matrix.) (**W**'s)

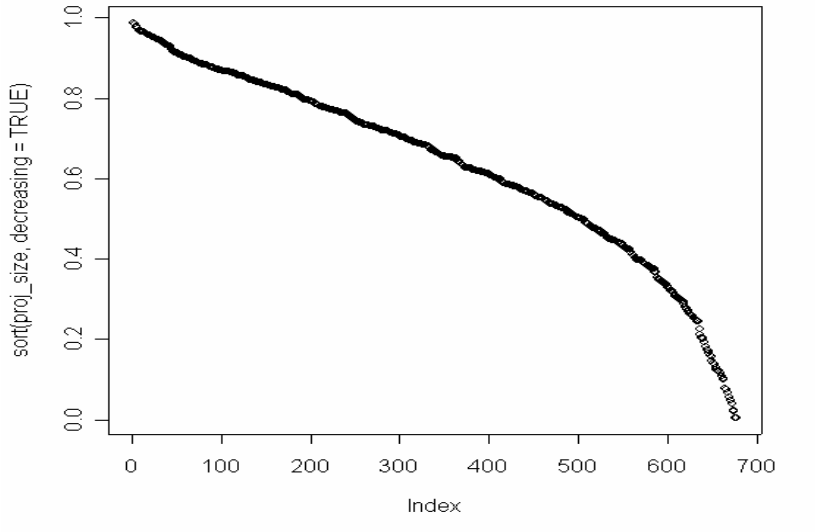
```
> yeast <- read.table(''proc_yeast_cycle.txt'',header=TRUE)
> yeastPCA <- prcomp(yeast,retx=TRUE,center=TRUE, scale.=FALSE,tol=NULL)
> summary(yeastPCA)
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7     PC8
Standard deviation     1.970 1.657 1.112 0.8849 0.6460 0.5841 0.5290 0.4939
Proportion of Variance 0.371 0.263 0.118 0.0749 0.0399 0.0326 0.0268 0.0233
Cumulative Proportion  0.371 0.634 0.752 0.8269 0.8668 0.8995 0.9262 0.9496
                          PC9   PC10   PC11      PC12
Standard deviation     0.4554 0.4133 0.3859 3.04e-06
Proportion of Variance 0.0198 0.0163 0.0142 0.00e+00
Cumulative Proportion  0.9694 0.9858 1.0000 1.00e+00

> plot(yeastPCA)
> plot(yeastPCA$x[,1],yeastPCA$x[,2])
```





```
> proj_size <-
(yeastPCA$x[,1]**2+yeastPCA$x[,2]**2)/rowSums(yeastPCA$x**2)
> plot(sort(proj_size,decreasing=TRUE))
```




```
> yeastPCA_corr <-
prcomp(yeast,retx=TRUE,center=TRUE,scale.=TRUE,tol=NULL)
> summary(yeastPCA_corr)
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7     PC8
Standard deviation     2.114 1.761 1.096 0.9576 0.7395 0.6659 0.6078 0.5728
Proportion of Variance 0.373 0.258 0.100 0.0764 0.0456 0.0370 0.0308 0.0273
Cumulative Proportion  0.373 0.631 0.731 0.8075 0.8531 0.8900 0.9208 0.9482
                          PC9   PC10   PC11     PC12
Standard deviation     0.5033 0.4465 0.4118 3.26e-06
Proportion of Variance 0.0211 0.0166 0.0141 0.00e+00
Cumulative Proportion  0.9693 0.9859 1.0000 1.00e+00
```

Et cetera...