Dimension Reduction (A):

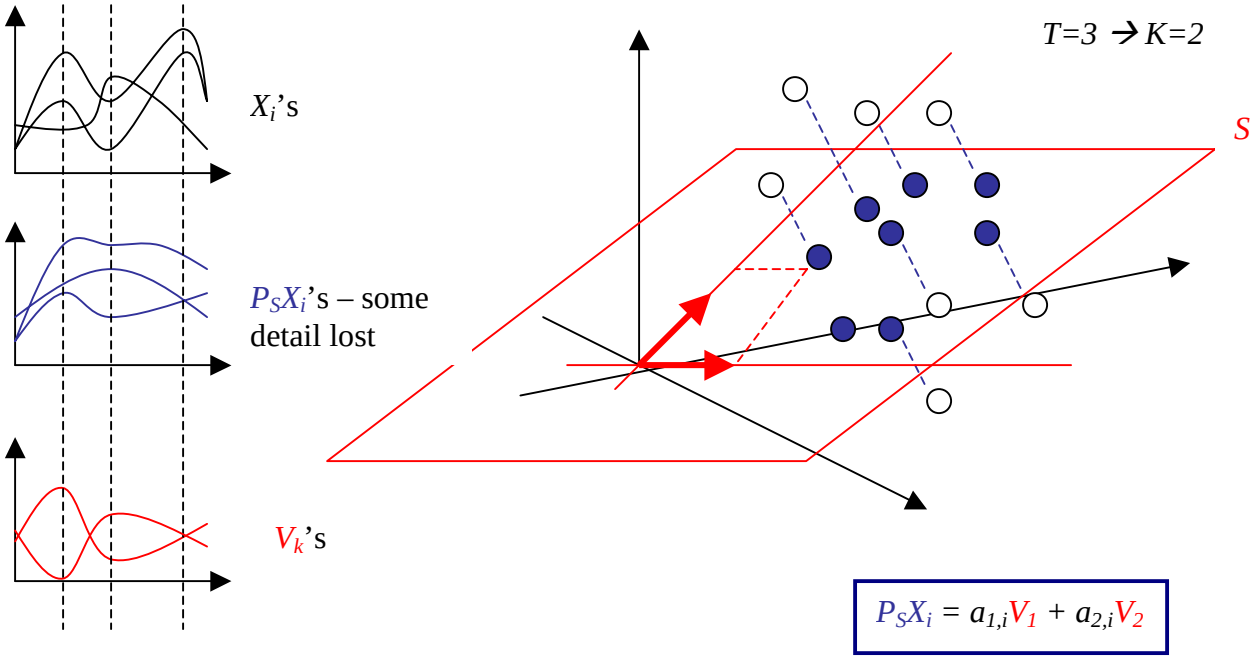
Think of the N rows of our data matrix as a cloud of points X_i , i=1...N in R^T.
Each individual expression profile corresponds to a point/vector in T dimensions.

Is there a way in which these profiles can be captured in lower dimension, substituting the original data cloud with its (orthogonal) projection on a subspace of R^T?

Neglecting temporarily the rationale by which this reduction can be achieved, suppose that the projection on a $K < T$ dimensional subspace S – e.g. K=2, a plane – provides a good representation of the data. The (orthogonal) projection of each individual profile on such subspace, $P_S X_i$, can be expressed as a linear combination of any collection of K linearly independent vectors constituting a basis of the subspace.

In particular, one selects an orthonormal basis: $\{V_1 \ldots V_K\}$, $||V_k|| = 1$, $V_k' V_l = 0$.

In the original coordinates, the basis vectors for the selected subspace are K "characteristic expression patterns". The reduction implies that each individual profile can, to a good approximation, be reconstructed additively from these K basic patterns.



Thus, thinking of dimension as a measure of complexity, if we can achieve a substantial dimension reduction we prove that the intrinsic complexity of the profiles is low.

Also, any further analysis can be restricted to the projected data (use of graphics, leaner computations during statistical analyses).

Dimension reduction is usually performed with an objective in mind:

What do we want to preserve?

Can we achieve a drop in dimension maintaining all or a large share of the information relative to a given feature of the data that is of interest to us?

For example, suppose that for each condition we observe an additional variable *Y*, together with gene expression (the *X*'s). *Y* could be:

Categorical (e.g. classification of the conditions)
Quantitative

We might want to study how *Y* depends on the *X*'s, and thus try to achieve a dimension reduction of the *X*'s that preserves information on *Y* contained in the original data. We will talk about this later.

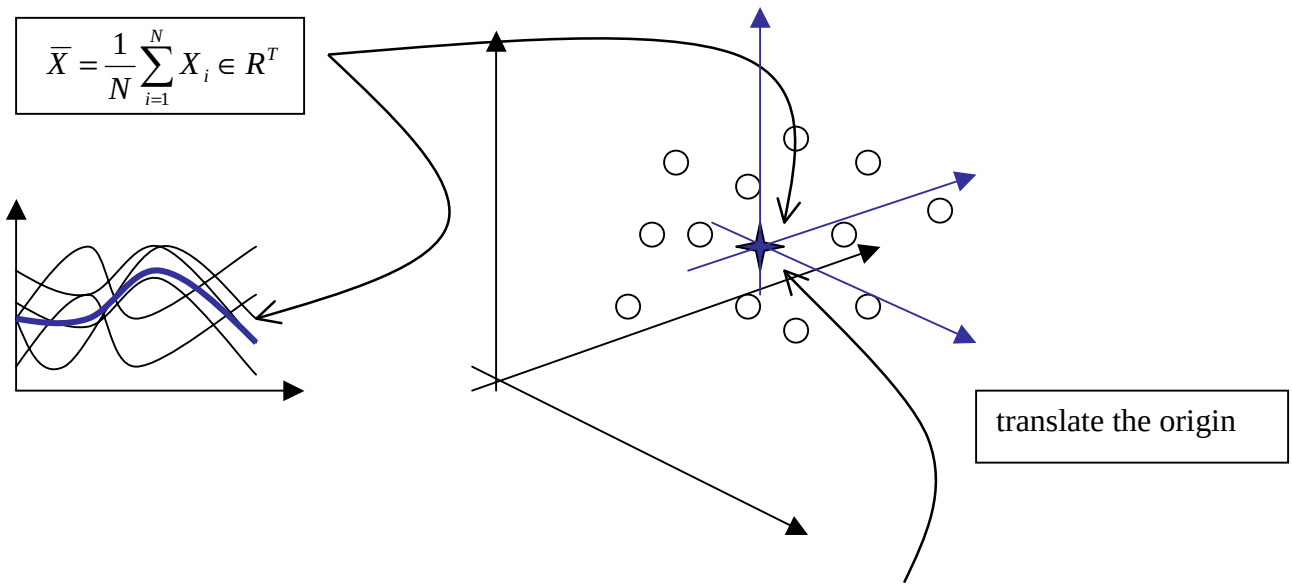
Here, we consider dimension reduction aimed at preserving the **structure of the *X*'s**…
Of course, what we mean by structure, or what aspect of the structure we are interested in, must be qualified! Conversely, rid of: unecessary detail? noise? artifacts?... (definitions?)

First dimension reduction method: PRINCIPAL COMPONENTS ANALYSIS

Structural feature of interest is the **variability of the *X*'s**

We look for a low-dimensional subspace capturing a large share of the overall variability of the data cloud.

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



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

$X_i, i = 1...N$ $\quad$ $(X_i - \overline{X}), i = 1...N$

On the other hand, whether we apply PCA to *X*'s that have been centered and standardized by row, or not, makes a difference… perfectly fine if the centering and standardization by row makes sense for our study, but we have to be aware of it!

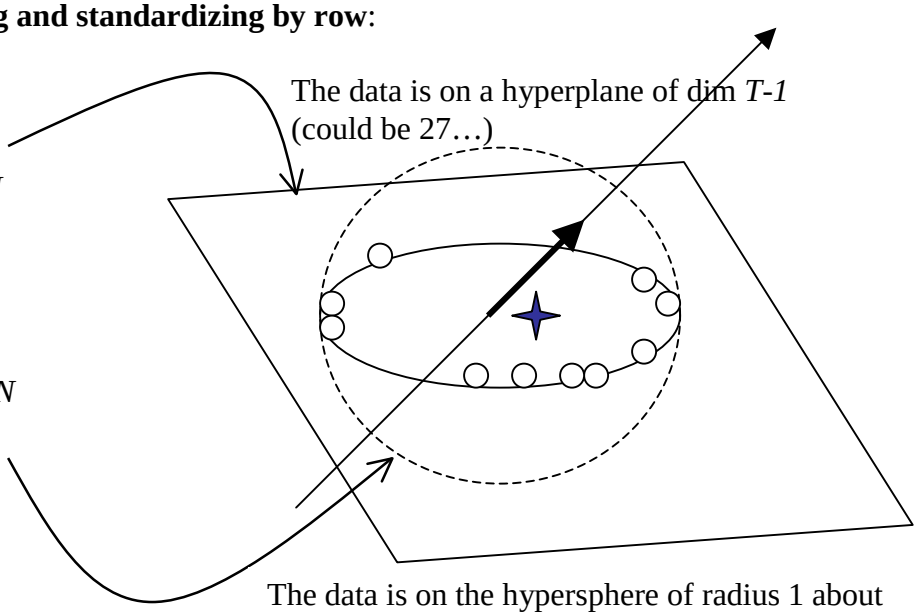
The geometry of centering and standardizing by row:

Centering by row

$$X_i '\vec{1} = \sum_{j=1}^{T} X_{i,j} = 0 \, , i = 1...N$$

Standardizing by row

$$\| X_i \|^2 = \sum_{j=1}^{T} X_{i,j}^2 = 1 \, , i = 1...N$$

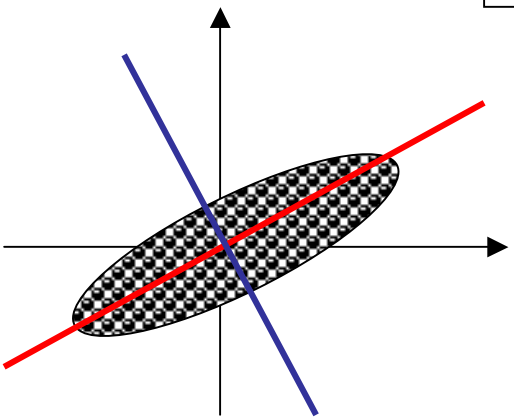


The **first phase of PCA** consist of determining a set of T orthogonal directions, ordered in terms of the variability displayed by the data along them.

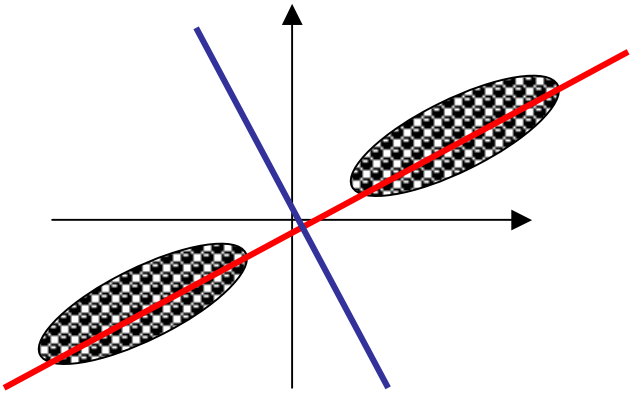
If the data cloud is (hyper) ellipsoidal, this is equivalent to determining the "natural axes" of the cloud, ordered in terms of their spread.

Aside: For Gaussian data, all there is to the structure is center (that we "translate out") and variability structure; no odd shapes, no clusters, no holes… thus PCA is in a sense an exhaustive dimension reduction tool. Not so for data whose structure is more complicated, but PCA can still be applied as a tool aiming at variability alone!

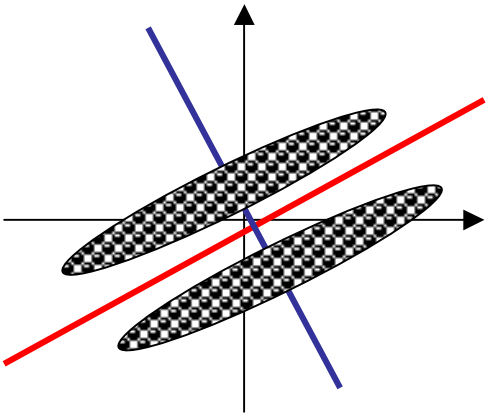
T=2



The first direction/component captures most of the variability



The first direction captures most of the variability, *and* the clustered structure



The first direction captures most of the variability, but *not* the clustered structure

Equivalently, consider the variance/covariance matrix over conditions as calculated on our expression profiles, with its spectral decomposition:

$$S = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})(X_i - \bar{X})' = \sum_{j=1}^{T}\lambda_j V_j V_j' \quad , \quad \lambda_1 \geq \ldots \geq \lambda_T \geq 0$$

eigenvectors

eigenvalues

Take the directions spanned by the eigenvectors, ordered in terms of the eigenvalue size

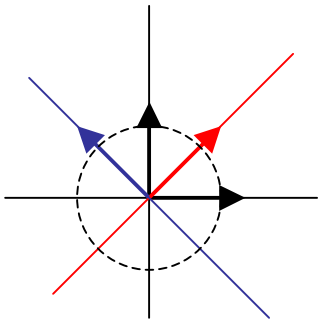
Eigenvalues are always real and non-negative, because a var/cov matrix is always (symm) non-negative definite. If one or more are 0, the data cloud lives in lower dimension to start with .

(*h* eigenvalues = 0, cloud lives in a *T-h* dimensional affine subspace; EXACTLY. For example at least one eigenvalue will be zero if the data were row-centered)

Eigenvectors $\{V_1 \ldots V_T\}$ are orthogonal $V_k'V_l = 0$
normal $||V_k|| = 1$

by construction; thus, they provide an orthonormal basis of R^T alternative to $\{e_1 \ldots e_T\}$.

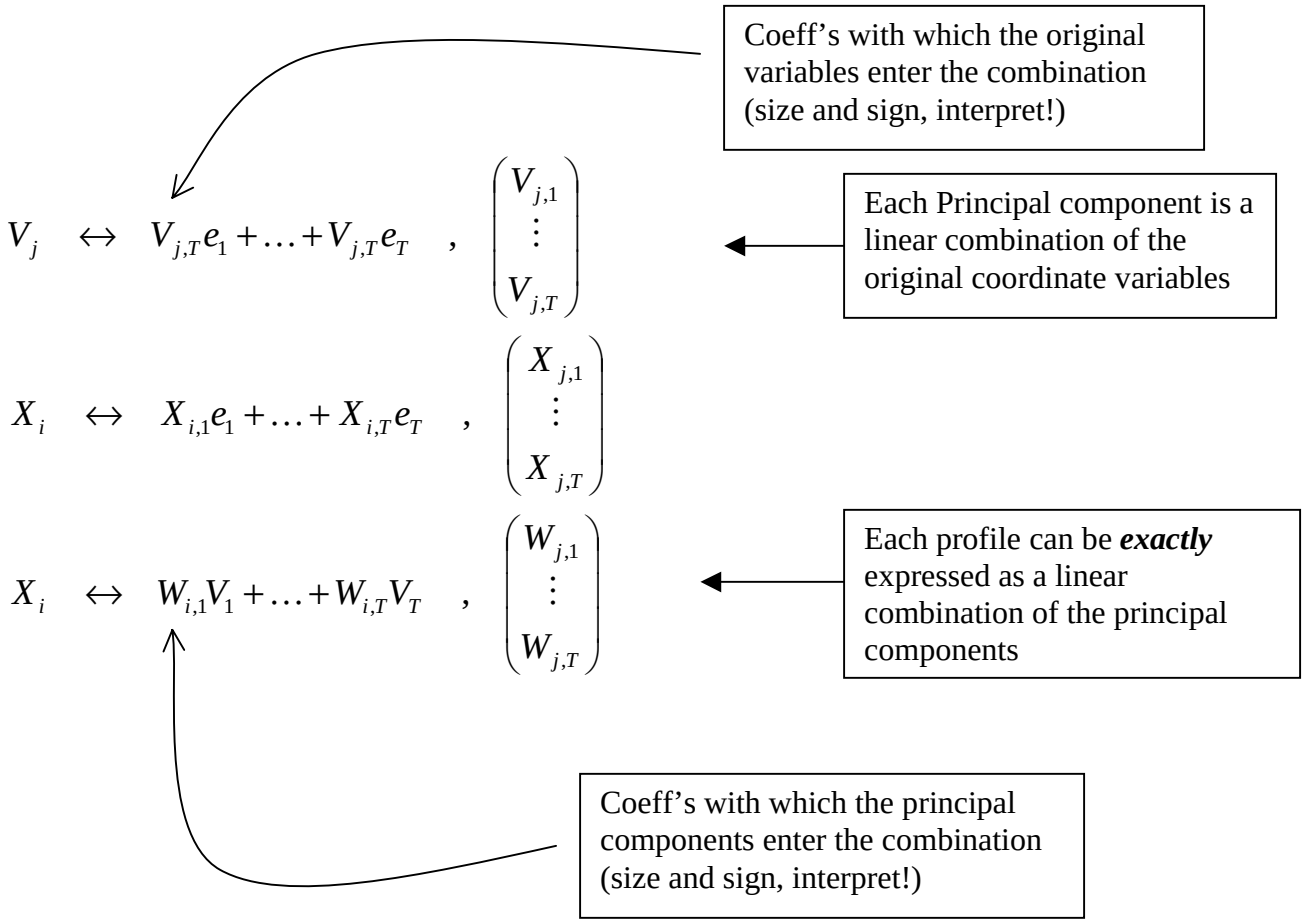
$\{V_1 \ldots V_T\}$ is a rotation of $\{e_1 \ldots e_T\}$.



Can also write

$$S = V\Lambda V', \quad V = (V_1 \ldots V_T), \quad \Lambda = \begin{pmatrix} \lambda_1 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \lambda_T \end{pmatrix}$$

orthogonal matrix (rotation)



$$W_{i,j} = V_j' X_i = P_{V_j} X_i$$

In the new coordinate system, the data cloud has a diagonal var/cov matrix

$$\frac{1}{N}\sum_{i=1}^{N}(W_i - \overline{W})(W_i - \overline{W})' = \Lambda$$

Passing from *N* profiles to *T* "characteristic patterns" through which all profiles can be exactly reconstructed, do we achieve a simplification/reduction?

Yes, but only to the extent that there isn't more than so much information in the data to start with…

… if we observe expression only on $T < N$ conditions, profiles are bound to be summarizable as combinations on *T* fundamental ones!

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

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

(at least in terms of variance structure…)

The **second phase of PCA** concerns how to determine how many components are necessary to achieve a good approximation.