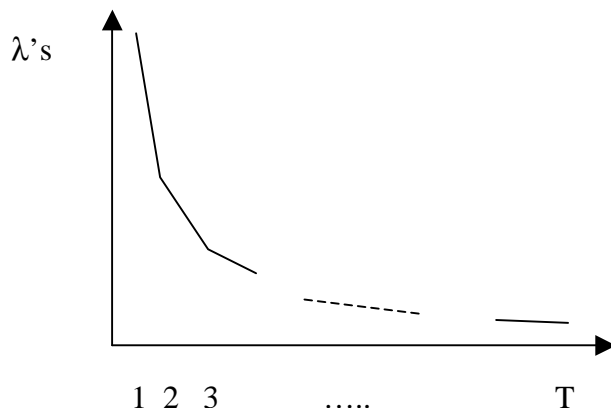


Dimension reduction (B):

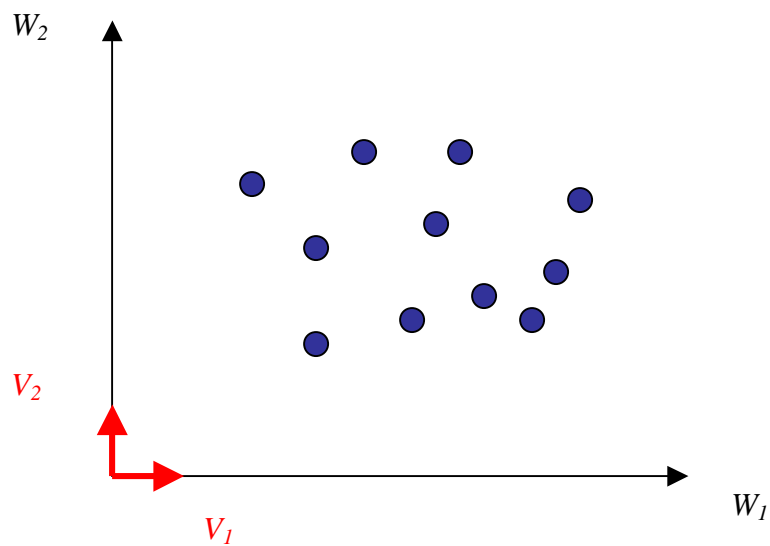
An important aside: PCA and plotting:

Scree plot: eigenvalues in non-increasing order



(visualizable) Projections:

Data cloud projected on the plane spanned by the first two principal components (this captures more variability than any other 2D projection of the cloud)





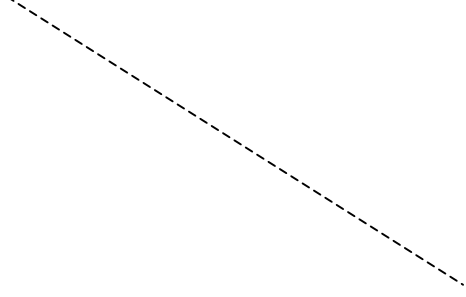

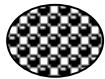


Or, data cloud projected on the space spanned by the first three principal components (this captures more variability than any other 3D projection of the cloud)

Cannot go further up maintaining a complete view... but

Scatter plot matrix:

2D “side views” of the T-dimensional cloud. This can be done with the original coordinates as well. When done with the principal components, the choice of side views is guided by variability explanation

W_1			
	W_2		
			
			W_T

Each plot in the matrix is usually rescaled to optimal “filling”

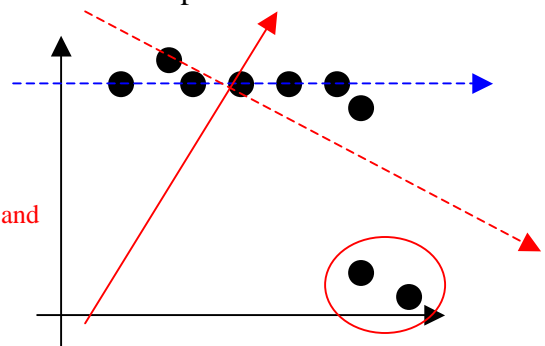
S is very sensitive to anomalous and/or “out of the bulk” points.

- In the 2D plot of the first two principal components, we might detect points that inflate var’s and/or distort cov’s (influential on the main variability directions).
- In the 2D plot of the last two principal components, we might detect points that induce directions that are not really there; mask singularity (influential on the small variability directions, and how many directions one chooses to retain).

Main variability axes may be different, and smaller variability axes may not exist, without these points. Need to identify and investigate them and attempt to understand if their anomalous position is due to errors.

- Try analysis with and without.
- Use a “robust version” of S .

Change of 1st direction, and 2nd direction, due to anomalous observations



Second phase of PCA: deciding how many directions to retain (K)

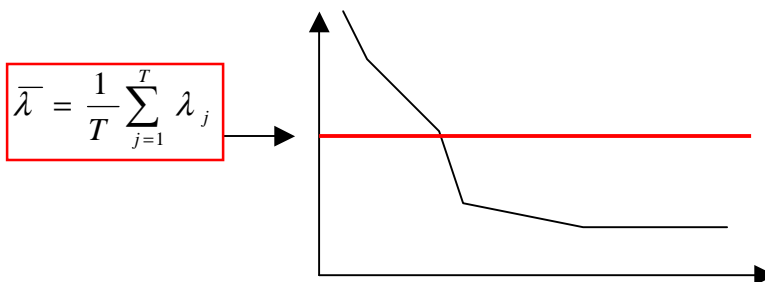
e.g. Is the first plane enough? We reason in terms of variability, and we'll decide based on it, analyzing the eigenvalues – whether it is enough with respect to other structural features is a different issue!

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

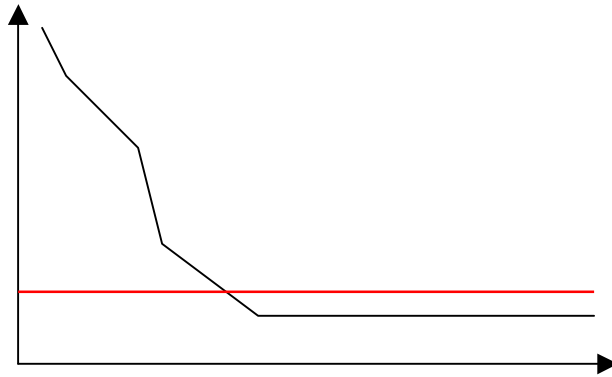
$$\frac{\lambda_1}{\sum_{j=1}^T \lambda_j} \quad \frac{\lambda_2}{\sum_{j=1}^T \lambda_j} \quad \dots \quad \frac{\lambda_r}{\sum_{j=1}^T \lambda_j} \quad \leftarrow \text{One by one}$$

$$\frac{\lambda_1}{\sum_{j=1}^T \lambda_j} \quad \frac{\lambda_1 + \lambda_2}{\sum_{j=1}^T \lambda_j} \quad \dots \quad \frac{\sum_{j=1}^T \lambda_j}{\sum_{j=1}^T \lambda_j} = 1 \quad \leftarrow \text{Cumulative. Stop when you reach, say, .80 i.e. 80\%}$$

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



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)

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

$$H_1 : \text{not}$$

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

More sophisticated methods exist – fairly large literature.

What are we missing?

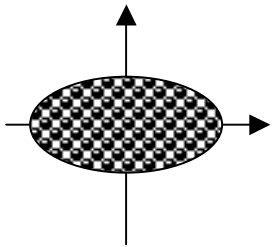
- With 1, we fix the percentage of variability we are missing
- With 2,3 and 4 we don't. It might not be negligible
- With 1 and 2, there might still be variability structure in the neglected directions
- With 3 and 4, the variability is approximately the same in all neglected directions

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.

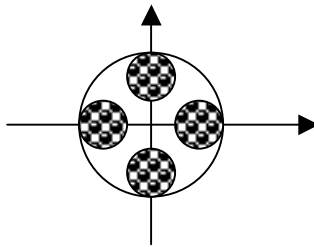
But if the percentage of variability we are missing is not very small (e.g. 20 or 30%), we ought to investigate what is going on there.

For example, if neglecting last two directions:

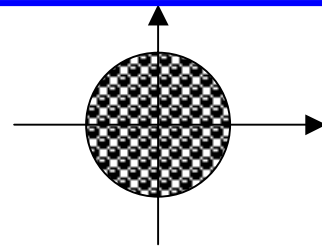
When it looks “structured”, is what we are throwing away “non-experimental” structure? Can we call it such if it plays out on small scale? Can we identify it as such even if it plays out on sizeable scale?



Still some variability structure



No variability structure... but some structure!



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

Of course if we are neglecting more than two or three directions a direct graphical investigation will not be possible.

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).

Interpreting principal components:

$$V_j = \sum_{l=1}^T V_{j,l} e_l$$

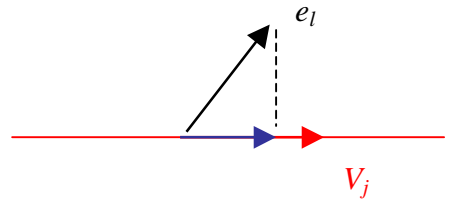
Contribution of the lth original coordinate (condition) to the jth component – sign and size

How close is e_l to V_j ?

$$P_{V_j} e_l = (V_j' e_l) V_j = (V_{j,l}) V_j$$

$$\| P_{V_j} e_l \|^2 = (V_j' e_l)^2 = (V_{j,l})^2$$

(recall norms are both = 1)

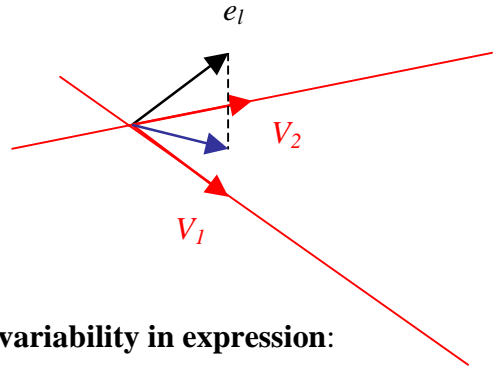


How close is e_l to $V_1 \dots V_K$ as a group, i.e. to $Span(V_1 \dots V_K)$?

$$P_{Span} e_l = \sum_{j=1}^K P_{V_j} e_l = \sum_{j=1}^K (V_j' e_l) V_j = \sum_{j=1}^K (V_{j,l}) V_j$$

$$\| P_{Span} e_l \|^2 = \sum_{j=1}^K \| P_{V_j} e_l \|^2 = \sum_{j=1}^K (V_{j,l})^2$$

Relatedly, can use first PCA component, or selected Subspace, to **rank (and select) original coordinates**.



i.e. **Conditions closest to the natural direction(s) of highest variability in expression:**
rank conditions in terms of square projection length.

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 ?

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

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

(recall one norm is =1)

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

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

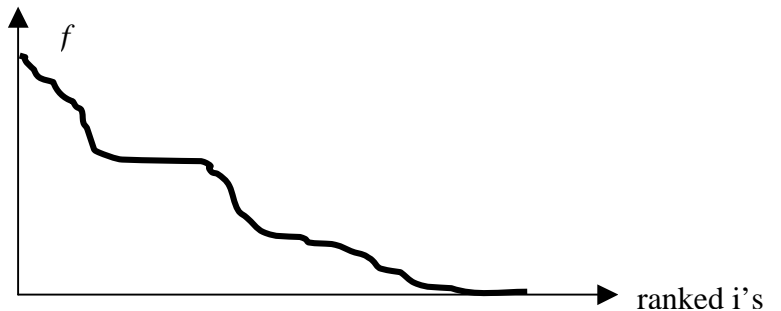
$$\frac{\|P_{\text{Span}} X_i\|^2}{\|X_i\|^2} = \frac{\sum_{j=1}^K \|P_{V_j} X_i\|^2}{\|X_i\|^2} = \frac{\sum_{j=1}^K (W_{i,j})^2}{\sum_{j=1}^T (W_{i,j})^2} = R_{X_i|V_1 \dots V_K}^2$$

Squared correlation between i th gene (actual row, profile) and j th PCA (synthetic row, basic pattern)

Determination coefficient from a linear ols fit (i th 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:**

rank genes in terms of $f_i = R_{X_i|V_1 \dots V_K}^2$.



Think about a “scrambling” scheme to provide a chance background for the ranking (“reference curves” on the ranking plot).

Using the correlation matrix instead of the covariance matrix:

Essential if the original coordinates represent quantities with different units of measure (not the case for us).

Also important if the original coordinates have very different variability sizes, and we do not want to take these differences under consideration... we are after “standardized variability structure”; that is, [correlation structure](#) (for us, think of a situation in which expression variability is much higher in some conditions than in others, and we do not consider these differences as relevant for our purposes).

$$S = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})(X_i - \bar{X})' = \begin{pmatrix} s_1^2 & & s_{1,T} \\ & \ddots & \\ s_{T,1} & & s_T^2 \end{pmatrix}$$

$$R = D(1/s_j) S D(1/s_j) = \frac{1}{N} \sum_{i=1}^N D(1/s_j)(X_i - \bar{X})(X_i - \bar{X})' D(1/s_j) = \begin{pmatrix} 1 & & r_{1,T} \\ & \ddots & \\ r_{T,1} & & 1 \end{pmatrix}$$

Correlation matrix, i.e. var/cov matrix of the standardized cloud $D(1/s_j)(X_i - \bar{X})$

...We are standardizing the data matrix by column.

The eigenstructures of S and R are different in both eigenvectors and eigenvalues.

Various:

A multivariate analysis text book: Gnanadesikan, *Methods for Statistical Data Analysis of Multivariate Observations*. Wiley.

[Singular Value Decomposition](#) (SVD) and PCA are equivalent analyses.

To think about as we read papers:

- How does “noise” look like?
- How do we construct a reference chance scenario?
- Under what circumstances can we claim that something “structured” we are throwing away is “non-experimental”? (“experimental artifacts”)
- What are the relationships between dimension reduction and clustering? (PCA is **NOT** a clustering technique).