How many components to keep?
- How to address the ambiguity of linear recombination?

Of course it depends on goal - denoising, data compression, visualization, modeling

Take PCA as an example to show basic strategies

**Recap**

$Y \overset{\text{f.t}}{=} XA$ | $Y$: $m \times r$, $X$: $m \times n$, $A$: $n \times r$

$Y \overset{\text{f.t}}{=} \underbrace{U^T}_{\text{"X"}} \underbrace{\Lambda V}_{\text{"A"}}$ | $U$: $n \times m$, $\Lambda$: $n \times n$ diag, $V$: $n \times r$

$X$: "sources"
$A$: "mixing matrix

rows of $U$ are left eigvecs of $YY^T$, orthonormal

$X$'s columns = eigvecs of $YY^T$, orthonormal

rows of $V$ are left eigvecs of $Y^TY$, orthonormal

$A = X^TY$, rows of $A$ are (left) eigvecs of $Y^TY$, orthogonal but not normalized

$$\Lambda = \begin{pmatrix} \sqrt{\lambda_1} & & \\ & \ddots & \\ & & \sqrt{\lambda_n} \end{pmatrix}$$

where $\lambda$'s are descending evals of $YY^T$ or $Y^TY$.

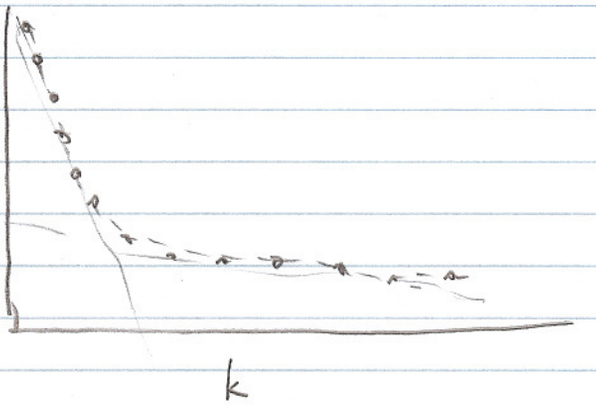
$YY^T$: $m \times m$, $Y^TY$: $r \times r$

If $n = \min(m, r)$, then $Y^{\text{f.t}} = Y$.

Typically, if $n < \min(m, r)$, then some variance is unexplained, namely, $\sum |Y^{\text{f.t}} - Y|^2 = \sum_{k=n+1 \text{ to } \min(m,r)} \lambda_k$

Several strategies to choose a smaller value of $n$ than $\min(m,n)$:

A "Scree plot"




guaranteed to descend ($\lambda$'s in order of importance); local slope indicates fraction of remaining variance explained by next component.

So choose the knee. (arbitrary).

OR, make such plots for random Y's, & choose the $\lambda$'s that are higher than expected for random data.

[Uncorrelated data, → expect $\lambda$'s all equal for (∞) vast amounts of data, but finite data will give a range of $\lambda$'s.]

[Can also generate surrogate correlated data based on some model of noise correlations]

[Eigenvalues of matrices of correlated Gaussians - a classic problem, many exact results but "difficult"]

B For data reduction/compression - choose some (arbitrary) fraction of variance to explain, (eg, 99%) & keep events up to that level

C. (Imaging analysis)

You may have some design variables, eg, $S_1, \cdots, S_D$, each as a time series $S_{d;m}$ - and you wish to "keep" the components that are correlated with or $S_d$, ie, fashion

keep $\lambda_h$ if $\left| \sum_{j=1}^{m} S_{d;j} X_{j,h} \right| >$ criterion  The retained $\lambda$'s might not be consecutive.

~ Fisher Disc, GIFA

D. (Visualization)
keep 2 or 3 $\lambda$'s

---

The analysis $Y^{fit} = XA$ is just as good a fit as

$$Y^{fit} = (XQ)(Q^{-1}A).$$

$Y^{fit} = U^T \Lambda V$ with $UU^T = I_{n \times n}$, $VV^T = I_{n \times n}$ discards to

bA only in terms of contribution to the variance.

Start with $Y^{fit} = XA$ and "move" normality from X to A:

$$Y^{fit} = (X\Lambda)(\Lambda^{-1}A), \text{ now } A' = \Lambda^{-1}A \text{ has orthogonal normalized rows.}$$

$A'_{jk}$ (or $A_{jk}$) indicates how much the $j^{th}$ factor (source) (column of X) contributes to the $k^{th}$ variable (column of Y)

Typically, the mixing matrix $A'$ will have most elements non-zero, i.e., most factors contribute to most observed variables

(22)

The "factor rotation" strategy is to find a rotation matrix $R$, for which $RA'$ has been "sparsified", i.e., has most elements either near 0, or near $\pm 1$.

Correspondence w Matlab: The TRANSPOSE of the above $A'$ is the "loading matrix" $A$, referred to in Matlab. $A_{MATLAB} = (A'_{JV})^T$.

VARIMAX criterion is to maximize (for $RA'$ composed of elements $a_{jk}$)

$$\sum_j \left( \frac{1}{r} \sum_k a_{jk}^4 - \gamma \left( \frac{1}{r} \sum_k a_{jk}^2 \right)^2 \right)$$

$\gamma = 1$ - VARIMAX

$\gamma = 0$ - "QUARTIMAX"

Since $A'$ is effectively $\Lambda^{-1} A$ & $\therefore$ has normalized rows, so $\gamma$-term is constant, independent of rotation.

The rationale here is that "simple" (sparse) mixing matrices are preferable - by making the mixing matrix as far from uniform as possible (i.e., as disperse as possible)

(23)

# ICA (Independent Components Analysis)

- another approach to the unmixing problem
- make the X's as dispersed as possible

"Cocktail Party Problem"-

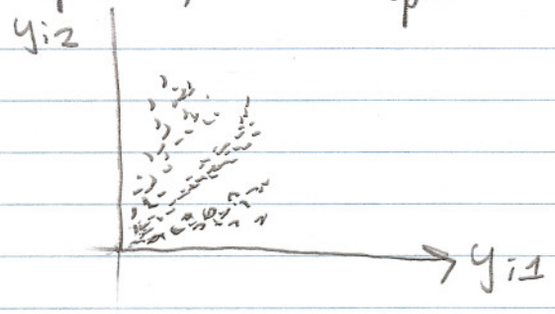
Each speaker $(1, \cdots, n)$ creates a time series $(1, \cdots, m)$, $X$ $(m \times n)$

Each microphone $(1 \cdots, n)$ picks up sound from each speaker @ efficacy $A$ $(n \times n)$.

Can we infer $X$ + $A$ from $Y = XA$? $(Y: m \times n)$

Above ambiguity problem would suggest that we can't — but there's another strategy we'll see.

3 speakers, 2 microphones - we can plot $y_{i1}$ vs $y_{i2}$




(Fig 1)

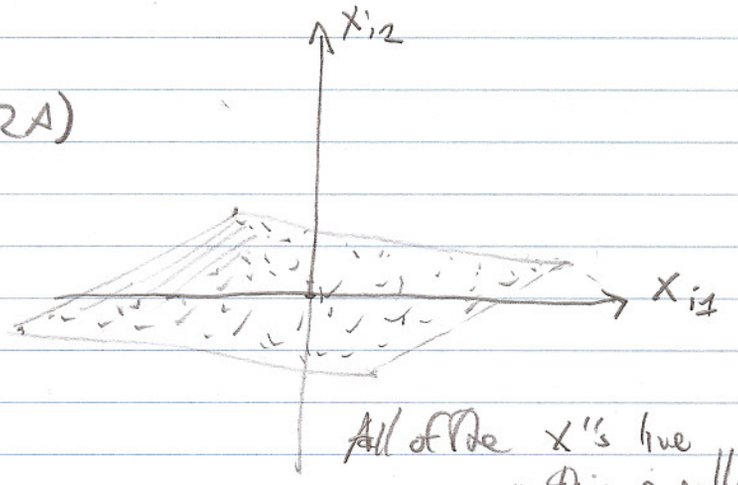
3 "plumes", corresponding to how each of the speakers is picked up by each microphone.

PCA will never identify this -- it can't find more than 2 components --

Another motivational example:

Say $Y$ is $m \times r$, $r \gg 2$, and you find an excellent fit of $Y_{fit} = XA$ for $X = m \times 2$.

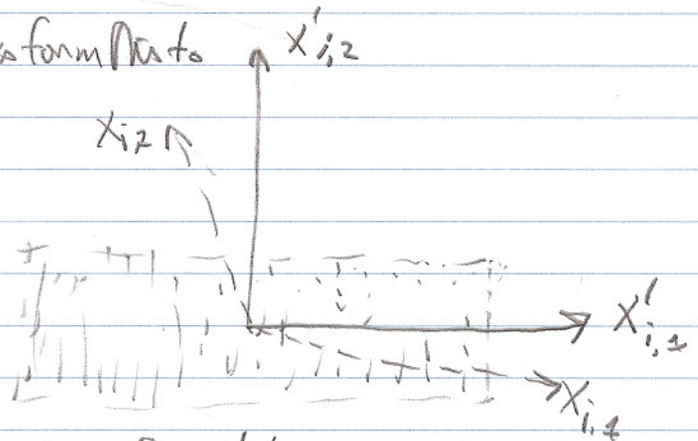
(Fig 2A)




All of the $x$'s live in this parallelogram.
$x_{i,1}$ explains most of the variance.

But we could transform data to

(Fig 2B)




In the new coords, the $x'$'s are independently distributed.

Common denominator in Fig 1, Fig 2 is that we made use of the observation that the points are not distributed in a Gaussian fashion.

(25).

The idea behind ICA is to find an $W$ for which

$WY$ has "unmixed" the observations $Y$ into components that are as non-Gaussian as possible.

See web tutorial by Aapo Hyvärinen

(Google ICA Tutorial)

All implementations are iterative (like VARIMAX)

One approach: In a Gaussian, if variance is $V$, then kurtosis

$$\left( K = \frac{\langle (x_j - \langle x_j \rangle)^4 \rangle - V^2}{3V^2} \right) \text{ is } 0.$$

$x_j$ = typical value of $j^{th}$ projection

So maximize $K^2$.

Another: Minimize the entropy of the projection $X$

Entropy of projection onto $x_j$ is

$$-\sum p_j(x) \log p_j(x)$$

Since a Gaussian has the maximum entropy possible, given a constraint on variance, minimizing the entropy (= maximizing the negentropy) is a reasonable def. of "non Gaussian".

Difficult (i.e. need lots of data) to estimate entropy from data, based on empirical $p_j(x)$

So - use a parametric approx to $p_j(x)$ ; this leads back to a moment criterion

Third approach Minimize the mutual information between the $X$'s. ("Independent Components")

(See Fig 2A,B).

Since total info in multivariate $Y$ is constant, minimizing the mutual info of the $X$'s is equivalent to maximizing their individual infos (= negentropy) → [second approach]

Essentially, ICA looks for non-Gaussian components.

So it will work really well for artifacts / some kinds - eye blinks EKG

- it is guaranteed to fail for Gaussian (or near-Gaussian) components.

---

Other general remarks:

- Biological components may not be independent
- None of the above make real use of dynamics -- same outcome if one time-shuffles the data

(? Fourier analysis first
? "Hierarchical Decomposition" Peppucci et al)