- basics of Linear Algebra
- PCA
- relationship of SVD to PCA

PCA Notes

Douglas Rubin

Applications of PCA - From data compression to feature reduction to more (will see applications in the lab)

Bases

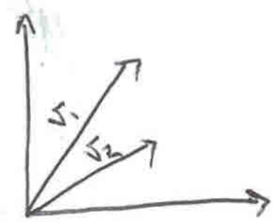
let V be a vector space. A set of vectors $\{v_1, \dots, v_n\}$, $v_i \in V$, is said to be a basis of V if:

1) they span the space (every vector in the space can be written as a linear combo. of these basis vectors)

2) they are linearly independent (ensures uniqueness of the components of the basis vectors)

examples in $\mathbb{R}^2$

1

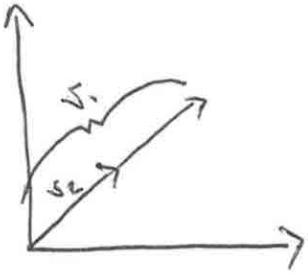


$$v_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

$$v_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

called "components" of the basis vectors

basis ✓ $\quad \begin{pmatrix} 3 \\ 4 \end{pmatrix} = 1 \begin{pmatrix} 1 \\ 2 \end{pmatrix} + 2 \begin{pmatrix} 1 \\ 1 \end{pmatrix}$

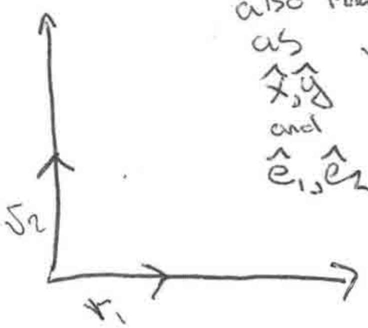
(2) 

$\cancel{\text{basis}}$ (doesn't span $\mathbb{R}^2$)

$$v_1 = \begin{pmatrix} 3 \\ 3 \end{pmatrix}$$

$$v_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$\begin{pmatrix} 3 \\ 4 \end{pmatrix} = \ldots$$

(3) 

basis ✓ (called the "standard" basis of $\mathbb{R}^2$)

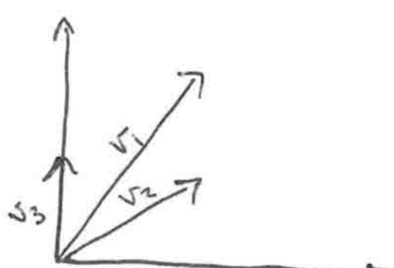
also known as $\hat{x}, \hat{y}$ and $\hat{e}_1, \hat{e}_2$

$$v_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

$$v_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

$$\begin{pmatrix} 3 \\ 4 \end{pmatrix} = 3\begin{pmatrix} 1 \\ 0 \end{pmatrix} + 4\begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

Known as an "orthonormal" basis since $\langle v_i, v_j \rangle = \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$
(i.e., v_1, v_2 are $\perp$, and each have length 1)

(4) 

$\cancel{\text{basis}}$ (not linearly indep., and hence not unique components)

$$v_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

$$v_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$v_3 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

$$\begin{pmatrix} 3 \\ 4 \end{pmatrix} = 1\begin{pmatrix} 1 \\ 2 \end{pmatrix} + 2\begin{pmatrix} 1 \\ 1 \end{pmatrix} + 0\begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

$$\begin{pmatrix} 3 \\ 4 \end{pmatrix} = 0\begin{pmatrix} 1 \\ 2 \end{pmatrix} + 3\begin{pmatrix} 1 \\ 1 \end{pmatrix} + 1\begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

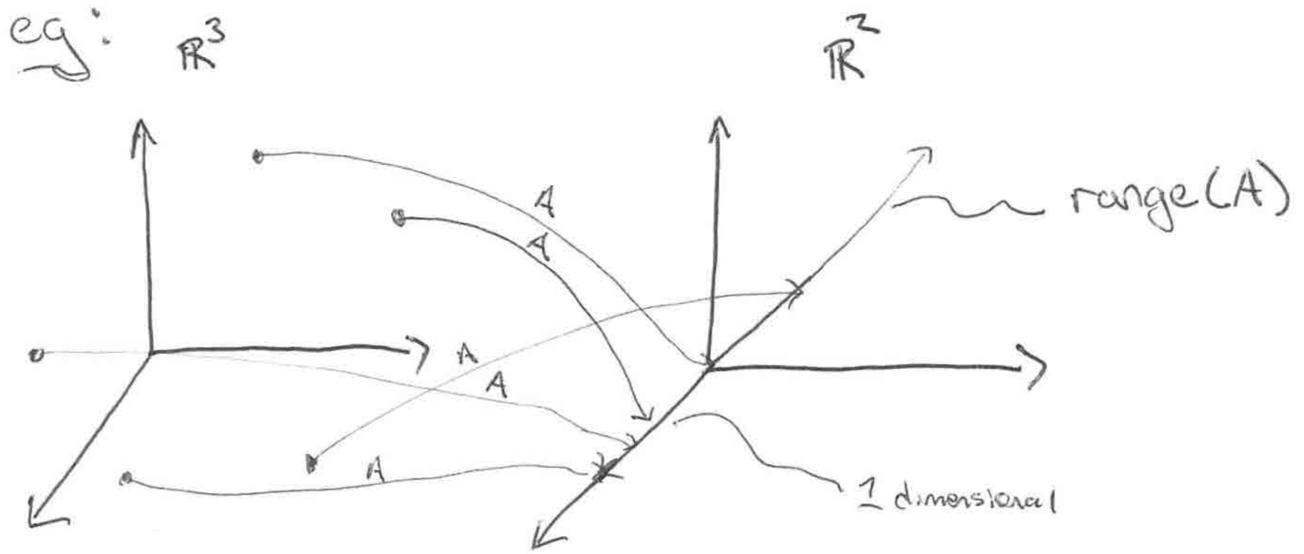
Matrix rank let $A \in \mathbb{R}^{m \times n}$, $x \in \mathbb{R}^n$

$$\underset{m \times n}{A} \; \underset{n \times 1}{x} = \underset{m \times 1}{\text{vector in } \mathbb{R}^m}$$, i.e., multiplication by A is a mapping from $\mathbb{R}^n$ to $\mathbb{R}^m$

$A : \mathbb{R}^n \to \mathbb{R}^m$

- mapping all vectors in $\mathbb{R}^n$ to $\mathbb{R}^m$ by multiplication of A results in a subspace in $\mathbb{R}^m$. (called range(A))

eg:



$A = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 1 & 2 \end{bmatrix}$ $\quad r = \dim \text{range} A = 1$

$r = \text{rank}(A) = \dim$ of resulting subspace $= \dim \text{range} A$

- one can prove that $\text{rank}(A) = \text{rank}(A^T)$

- Also, $\text{rank}(A) = \#$ linearly indep. columns of A
  $\text{rank}(A^T) = \#$ " " rows " A

Eigenvectors / Eigenvalues

let $A \in \mathbb{R}^{n \times n}$. The eigenvectors of A are all vectors, $v \in \mathbb{R}^n$, such that $Av = \lambda v$, $(v \neq \vec{0})$, where $\lambda \in \mathbb{R}$ and is called the corresponding eigenvalue.

- thus, eigenvectors only have their length changed (and not orientation) upon application of A.

eg: $A = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix} \Rightarrow v_1 = \begin{pmatrix} 1 \\ -1 \end{pmatrix},\ \lambda_1 = -1$

$v_2 = \begin{pmatrix} 1 \\ -2 \end{pmatrix},\ \lambda_2 = -2$

$$\begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix}\begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \end{bmatrix} = -1\begin{bmatrix} 1 \\ -1 \end{bmatrix} \checkmark$$

$$\begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix}\begin{bmatrix} 1 \\ -2 \end{bmatrix} = \begin{bmatrix} -2 \\ 4 \end{bmatrix} = -2\begin{bmatrix} 1 \\ -2 \end{bmatrix} \checkmark$$

Change of bases of a Matrix

let $A \in \mathbb{R}^{n \times n}$, and assume that A is expressed in the standard basis. Then it can be shown that the matrix expressed in a new basis $\{v_1, \dots, v_n\}$, B, is given by:

$B = CAC^{-1}$, where the columns of C are $\{v_1, \dots, v_n\}$ $(C = [v_1, \dots, v_n])$

- the eigenvectors of a matrix are a very nice basis in which to re-express the matrix, since they add many 0s to the new matrix, and hence make it more simple
- in particular, if A has n linearly indep. eigenvectors, then we can "diagonalize" A by choosing the eigenvectors as the new basis:

$$B = C^{-1}AC = C^{-1}A[v_1, \dots, v_n] = C^{-1}[Av_1, \dots, Av_n]$$

$$= C^{-1}[\lambda_1 v_1, \dots, \lambda_n v_n] = [\lambda_1 C^{-1}v_1, \dots, \lambda_n C^{-1}v_n]$$

Claim: $C^{-1}v_1 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$; $C^{-1}v_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ \vdots \end{bmatrix}$ $\dots$ $C^{-1}v_n = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}$

if true, then $v_1 = C\begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} | & & | \\ v_1, & \dots, & v_n \\ | & & | \end{bmatrix}\begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = 1v_1 + 0v_2 + \dots + 0v_n = v_1$ ✓

$$\Rightarrow B = [\lambda_1 e_1, \dots \lambda_n e_n] = \begin{bmatrix} \lambda_1 & & 0 \\ & \lambda_2 & \\ 0 & & \ddots & \lambda_n \end{bmatrix}$$

Spectral Thereom (For real, symmetric matrices)

- one of the most important thereoms in linear algebra

let $A \in \mathbb{R}^{n\times n}$ be a real, symmetric matrix ($A^T = A$), then the eigenvectors of A form an orthonormal basis of $\mathbb{R}^n$, and the corresponding eigen values are real and positive.

- An amazing result because a matrix need not have n distinct eigenvectors, let alone n distinct eigenvectors that are orthonormal.

eg: $A = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} \Rightarrow$ $v_1 = \begin{bmatrix} 1/\sqrt{5} \\ 2/\sqrt{5} \end{bmatrix}, \lambda_1 = 5$

$v_2 = \begin{bmatrix} -2/\sqrt{5} \\ 1/\sqrt{5} \end{bmatrix}, \lambda_2 = 0$

$v_1 \cdot v_1 = \frac{1}{5} + \frac{4}{5} = 1$ ✓

$v_2 \cdot v_2 = \frac{4}{5} + \frac{1}{5} = 1$ ✓

$v_1 \cdot v_2 = \frac{-2}{5} + \frac{2}{5} = 0$ ✓

Corollary

Any symmetric matrix $A \in \mathbb{R}^{n \times n}$ can be diagonalized by a basis of its eigenvectors:

- this follows from the spectral thereom and the previous section on change of bases

eg: $A = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$

$$B = \begin{bmatrix} \frac{1}{\sqrt{5}} & -2/\sqrt{5} \\ 2/\sqrt{5} & 1/\sqrt{5} \end{bmatrix}^{-1} \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} \begin{bmatrix} 1/\sqrt{5} & -2/\sqrt{5} \\ 2/\sqrt{5} & 1/\sqrt{5} \end{bmatrix}$$

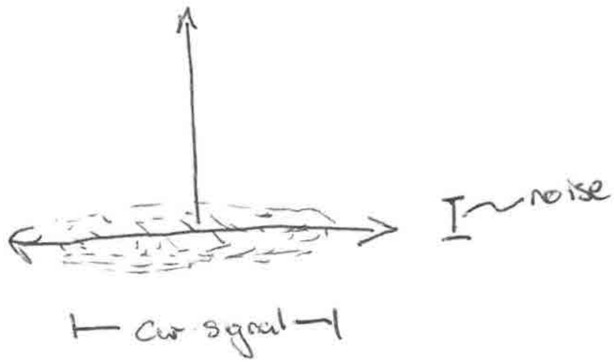
$$= \begin{bmatrix} 5 & 0 \\ 0 & 0 \end{bmatrix}$$

PCA

- interpretation of data can be complicated by noise and redundancy (correlation of the measured variables) in the data

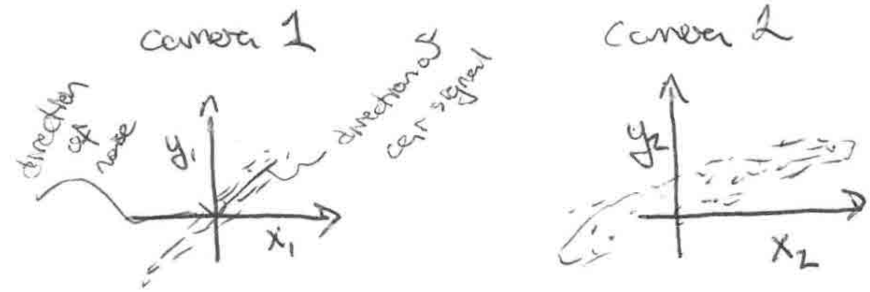
- consider a car oscillating back + forth on a spring w/ 2 cameras at arbitrary angles measuring it's position as a function of time.

- intuitively, we know that the relevant signal is contained along the X-axis, and thus only 1 camera is needed and should be position along the X-axis.



in fact, we would like to eliminate the noise and just measure x-positions.

- what we actually measure is noise + correlation in the data:



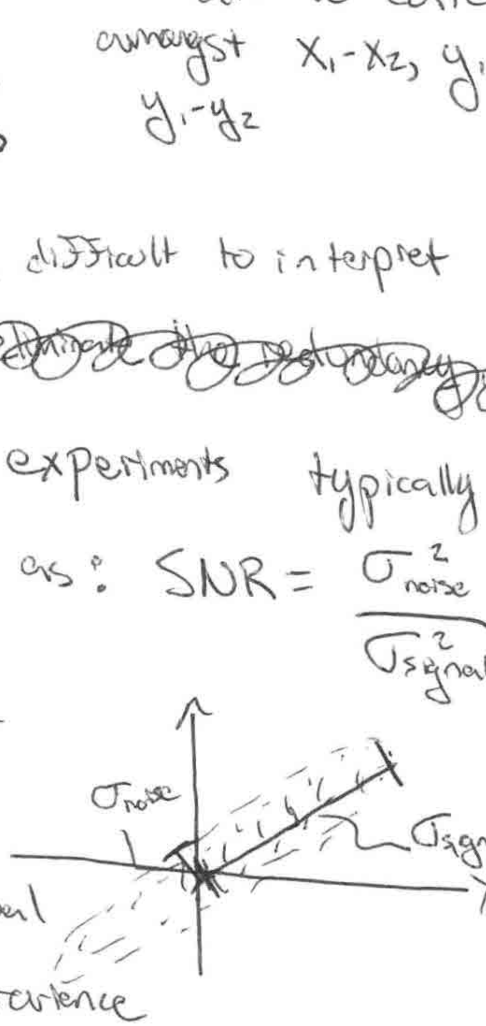
- will also be correlation amongst x_1-x_2, y_1-x_2, y_2-x_1, y_1-y_2

- the noise and redundancy make the data difficult to interpret

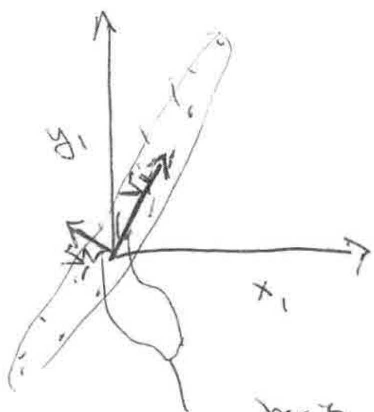
noise

- quality of measurements in most experiments typically measured by signal-to-noise-ratio as: $SNR = \frac{\sigma^2_{noise}}{\sigma^2_{signal}}$ (ie, ratio of variances is a typical way to measure noise)

- it is natural to expect that, for any decently designed experiment, directions w/ higher variation/variance correspond to actual signal, while directions w/ lower variance are probably associated more w/ noise.



redundancy / correlation



- Features x_1, y_1 are redundent (and therefore confuse interpretation of the data)
- that is, we did not originally choose the best basis w/ which to express our data.

new basis
v_1 = 1st principle component
v_2 = 2nd principle component

PCA Solution

Choose a new basis st. we can
(1) reduce redundency / correlation
(2) eliminate noise-y Features under the assumption that directions w/ the smallest verience probably correspond more to noise

the data above re-expressed in the new basis looks like: PCA would just choose the 1st principle component here.

- Diagonalizing the covariance Matrix
  - We can in fact eliminate all correlation if we can choose ~~diago~~ a basis which diagonalizes the covariance matrix

- let $X \in \mathbb{R}^{m \times n}$ be our design matrix, and let's assume it has been centered by its mean (i.e., subtract off the mean from each column) (since most of the math in the rest of these notes requires that the data lie approximately in a subspace, this assumption will be necessary since any valid subspace needs $\vec{0}$, the zero vector). Then:

$$\Sigma_{ij} = \frac{1}{m-1} \underbrace{\sum_{k=1}^{m}}_{\text{~~sums~~ sum over rows}} (X_{ki} - \cancel{\mu_i}^{\,0})(X_{kj} - \cancel{\mu_j}^{\,0}) = \frac{1}{m-1} \sum_{k=1}^{m} (X^T)_{ik} X_{kj}$$

$$= \frac{1}{m-1} (X^T X)_{ij}$$

$$\Rightarrow \underset{n \times n}{\Sigma} = \frac{1}{m-1} \underset{n \times m}{X^T} \underset{m \times n}{X}$$

- notice that Σ is symmetric:

$$\Sigma^T = \left(\frac{1}{m-1} X^T X\right)^T = \frac{1}{m-1} X^T X^{T^T} = \frac{1}{m-1} X^T X = \Sigma,$$

and it can therefore from the spectral theorem be diagonalized by an orthonormal basis of its eigenvectors. let $\{v_1, \dots, v_n\}$ be the eigenvectors of Σ. Then:

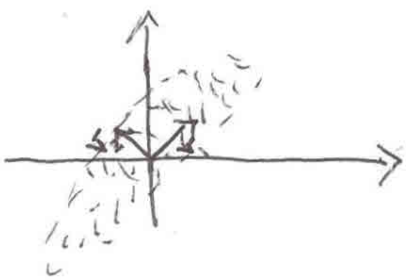
note that $\sigma_i^2 = \lambda_i \geq 0$ by the spectral theorem (as it should be)

$$S = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_n^2 \end{bmatrix} = [v_1, \dots, v_n]^T \Sigma [v_1, \dots, v_n]$$

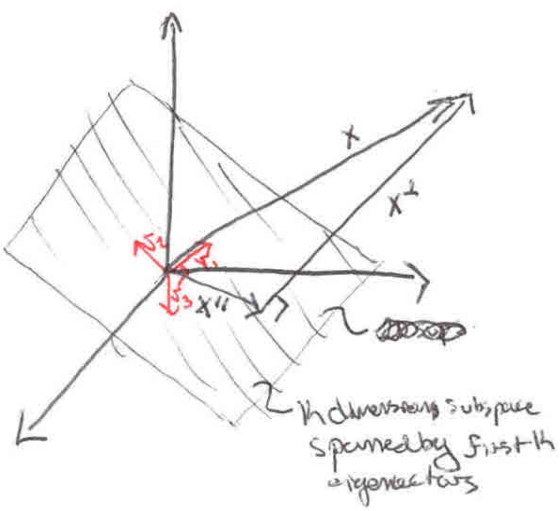
covariance matrix in new basis

eigenvalues

relabelling the λs w/ σ^2s since they are the variances along the $v_1, \dots, v_n$ directions where we have specifically arranged the eigenvectors / eigenvalues st: $\sigma_1^2 \geq \sigma_2^2 \geq \ldots \geq \sigma_n^2$. Thus, the change of basis corresponds to the new basis vectors in the figure, since they eliminate all correlation. The $\sigma_1, \ldots, \sigma_n$ are thus the variances in those directions.



reducing dimensionality w/ PCA



To reduce the dimensionality of the data, while keeping the components which correspond most to actual signal, and while eliminating corellation, we thus project X onto the subspace created by the 1st k eigenvectors:

$$X = X^{\parallel} + X^{\perp} = \alpha_1 v_1 + \alpha_2 v_2 + \ldots + \alpha_k v_k + \alpha_{k+1} v_{k+1} + \ldots + \alpha_n v_n$$

- since $\{v_1, \ldots, v_n\}$ is orthonormal:

$$\langle X, v_j \rangle = \langle \alpha_1 v_1 + \ldots + \alpha_n v_n, v_j \rangle = \alpha_j \langle v_j, v_j \rangle = \alpha_j$$

$$\Rightarrow \quad X = \underbrace{\langle X, v_1 \rangle v_1 + \ldots + \langle X, v_k \rangle v_k}_{X^{\parallel}} + \underbrace{\langle X, v_{k+1} \rangle v_{k+1} + \ldots + \langle X, v_n \rangle v_n}_{X^{\perp}}$$

$\longrightarrow$

thus, by reducing the # features w/ PCA to k ($<n$), the new design matrix is:

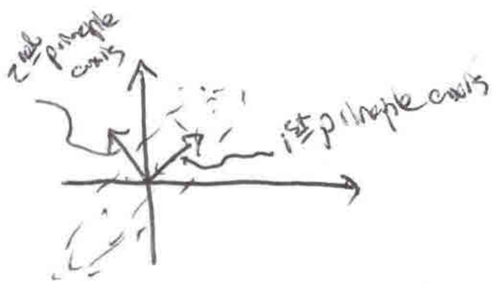
$$X_{pca} = \begin{bmatrix} — & x^{(1)\prime\prime} & — \\ & \vdots & \\ — & x^{(m)\prime\prime} & — \end{bmatrix} = \underbrace{\begin{bmatrix} \langle x^{(1)}, v_1 \rangle & \langle x^{(1)}, v_2 \rangle & \dots & \langle x^{(1)}, v_k \rangle \\ \vdots & & & \\ \langle x^{(m)}, v_1 \rangle & \langle x^{(m)}, v_2 \rangle & \dots & \langle x^{(m)}, v_k \rangle \end{bmatrix}}_{m \times k}$$

$$= \underset{m\times n}{X} \underset{n \times k}{[v_1, \dots, v_k]}$$

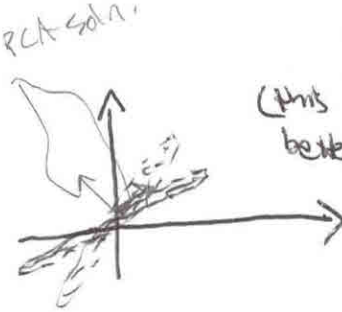
Assumptions made under PCA:

① principle directions are orthogonal to each other (necessary in order to diagonalize Σ w/ its eigenvectors)

i.e. data looks like:

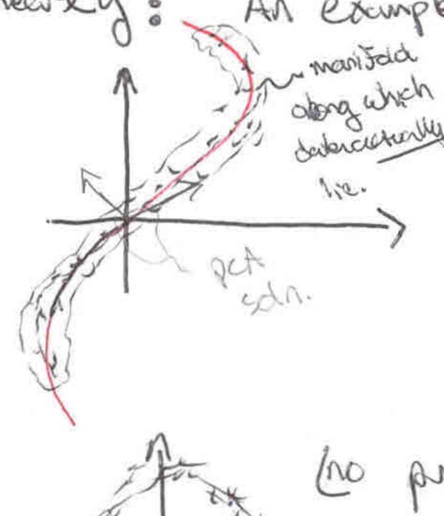


and not (for eg)



(this would be better modeled w/ ICA)

② data behave linearly: An example of non-linearity would be



(data here would probably be modelled better by some manifold learning method like kernel PCA)

or even worse

(no preferred direction)

→

- thus, putting these both together, the data should be distributed more or less like ellipsoidal blobs, i.e., like a Gaussian.

(3) directions w/ higher variance correspond to meaningful signal dynamics, while directions w/ lower variance correspond to noise.

Other interpretations of PCA

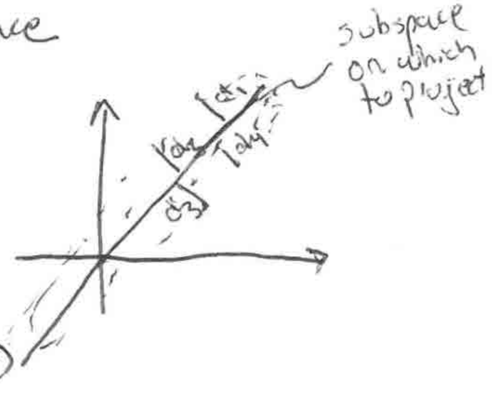
diagonalizing Σ and keeping k components

1) re-expressing basis to eliminate correlation and projecting in a low dim. subspace w/ the most signal

due to SVD

2) ~~maximizing~~ ~~[illegible]~~
Finding a subspace which minimizes sum of squared distances ~~to~~ of data to sub-space (similar to linear regression)

- Find $v_1, \dots, v_k$ which span the subspace which minimizes
$$\sum_{i=1}^{m} d_i^2$$



3) ~~Finding a low dimensional subspace which is the range of rank k (< min(m,n)) the low rank k matrix $\tilde{X}$, that is as "close" as possible to X according to the Frobenius norm: $\tilde{X} = \arg\min_{\tilde{X} \in \mathbb{R}^{m \times n},\ \mathrm{rank}\,\tilde{X} = k} \|X - \tilde{X}\|_F^2$~~

See end of notes for 3rd interpretation

(14)

ve
ro
,VD
vel the definition of the right singular vectors)

4) Finding direction w/ maximal variance, finding direction w/ maximal variance and $\perp$ to preceeding directions,

5) Finding a "compression matrix" $W \in \mathbb{R}^{k \times n}$ and "recovery" . . . matrix $U \in \mathbb{R}^{n \times k}$ which minimizes the recovery error induced by a compression to a lower dim. space via a matrix multiplication:

$$W, U = \underset{W \in \mathbb{R}^{k \times n},\, U \in \mathbb{R}^{n \times k}}{\operatorname{argmin}} \sum_{i=1}^{m} \| x^{(i)} - U W x^{(i)} \|^2$$

SVD and its connection to PCA

let $X \in \mathbb{R}^{m \times n}$ be any real matrix w/ rank$=r$, then the SVD decomposition of X is:

$X = UDV^T$; where U, V are orthogonal matrices, and D is a diagonal matrix

3 different types

σs arranged in descending order

$$)\, X = \underset{m \times m}{\begin{bmatrix} | & & | \\ u_1 & \cdots & u_m \\ | & & | \end{bmatrix}} \underset{m \times n}{\begin{bmatrix} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_r & \\ 0 & & & 0 \\ & & & \ddots \\ & & & & 0 \end{bmatrix}} \underset{n \times n}{\begin{bmatrix} - & v_1^T & - \\ & \vdots & \\ - & v_n^T & - \end{bmatrix}}$$

- us called left singular vectors
- σs called singular values
- vs called right singular vectors

$$2) \quad = \underset{m \times r}{\begin{bmatrix} | & & | \\ u_1 & \cdots & u_r \\ | & & | \end{bmatrix}} \underset{r \times r}{\begin{bmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_r \end{bmatrix}} \underset{r \times n}{\begin{bmatrix} - & v_1^T & - \\ & \vdots & \\ - & v_r^T & - \end{bmatrix}}$$

③ $\tilde{X}$ = best rank k approx. to X = $\underbrace{\begin{bmatrix} | & & | \\ u_1 & \cdots & u_k \\ | & & | \end{bmatrix}}_{m \times k} \underbrace{\begin{bmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_k \end{bmatrix}}_{k \times k} \underbrace{\begin{bmatrix} - v_1^T - \\ \vdots \\ - v_k^T - \end{bmatrix}}_{k \times n}$ $(k<r)$

(called "truncated" SVD)

- one can prove that the right singular vectors are the eigenvectors of X^TX and the singular values are the square roots of the corresponding eigenvalues.

- For PCA, the vs and σs are typically computed w/ SVD since it is more numerically stable, and since calculating the eigenvectors of $\Sigma \in \mathbb{R}^{n \times n}$ can be daunting for $n \gg 1$.

- in summary, SVD can provide the eigenvectors / eigenvalues needed for PCA, and it provides a few additional ways of interpreting what PCA does.

---

note to SVD

③ For a design matrix $X \in \mathbb{R}^{m \times n}$ w/ rank r, finding a lower rank $(k<r)$ matrix, $\tilde{X}$ that is as close as possible to X according to the Frobenius norm: $\tilde{X} = \underset{\substack{\tilde{X} \in \mathbb{R}^{m \times n}: \\ \text{rank}(\tilde{X})=k}}{\arg\min} \|X - \tilde{X}\|_F^2$.

- ~~note that the subspace of $\mathbb{R}^n$ spanned by the rows of $\tilde{X}$ (i.e. the range of $\tilde{X}^T$) corresponds to the space spanned by the k dim. PCA approximation~~

- note that $\tilde{X}$ is written in the original basis. If we write $\tilde{X}$ in the PCA basis, the new matrix would be exactly X_{PCA}.