The Singular Value Decomposition

Recall: If A is $n \times n$, and has n lin. ind. eigenvectors $\vec{v}_1, \ldots, \vec{v}_n$ with $\|\vec{v}_i\| = 1$, the A is <u>diagonalizable</u>

$$A = P D P^{-1}$$

$$P = (\vec{v}_1\ \vec{v}_2 \cdots \vec{v}_n) \qquad D = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}$$

What if A is $m \times n$ (not square)? Or <u>not</u> diagonalizable?

In this case, the SVD tells you everything you need to know about the matrix.

$$A = U S V^T$$

- A: $m \times n$ matrix
- U: $m \times m$ orthogonal
- S: diagonal, $m \times n$, $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k$
- V^T: $n \times n$ orthogonal

$$= \text{Rotation/reflection} \circ \text{Scaling} \circ \text{Rotation/reflection}$$

<u>Naïve calculation</u> (assume A real)

1. Form $B = A^T A$
2. Find eigenvalues/vectors of B: $\vec{v}_1 \cdots \vec{v}_n$, $\lambda_1 \cdots \lambda_n$

   $\Rightarrow B = V \Lambda V^T$
3. Set $U = A V (\Lambda)^{-1/2}$ $\quad \left(U^T U = (\Lambda)^{-1/2} V^T A^T A V (\Lambda)^{-1/2} = I \right)$
4. $\Rightarrow A = U \underbrace{(\Lambda)^{1/2}}_{S} V^T$

If $m > n$ and A has rank k, then

$$A = U_{m \times k} \, S_{k \times k} \underbrace{V^T}_{k \times n}$$

The pseudo-inverse of A is defined to be:

$$\underbrace{A^+}_{\text{A dagger}} = V S^{-1} U^T$$

(also called Moore-Penrose inverse)

And even though A is not invertible,

$$A^+ A = (V S^{-1} U^T)(U S V^T)$$

$$= V S^{-1} U^T U S V^T$$

$$= V V^T = I \text{ when } V \text{ is square}$$

= projection onto row-space when rank $A < n$.

Lastly the SVD offers the best 2-norm (rank k) matrix approximation:

Let A be $m \times n$ with rank r. Let $k < r$.

Then $\underset{\substack{A', \\ \text{rank } A' = k}}{\arg\min} \| A - A' \|_2 = \underbrace{U \, S_k \, V^T}_{\text{truncated SVD}} = (m \times n) \cdot \begin{pmatrix} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_k & \\ & & & 0 \ddots 0 \end{pmatrix} (n \times n)$

$n \times n$

$$\| A - U_k S_k V_k^T \|_2 = \| U S V^T - U S_k V^T \|_2$$

$$= \| U (S - S_k) V^T \|_2$$

$$= \| S - S_k \|_2 = \boxed{\sigma_{k+1}}.$$

All methods are some form of dimension reduction, etc.

Principal Component Analysis

Problem setup (in the continuous case)

Two main tenets:
- variation in data provides information
- correlation between data reduces amount of information.

(i.e. in the normal case, zero corr $\iff$ indep.)

Idea: Transform (linearly) m variates into $k < m$ variates that contain *almost* as much variation as the original, but are *approximately* mutually independent.

Eg. $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{matrix} aZ_1 + bZ_2 \\ cZ_1 + dZ_2 \end{matrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix}$ $\quad Z_1, Z_2 \sim N(0,1)$

If we know $\mathrm{Cov}(X_i, X_j)$ how could we decouple X_1, X_2 into Z_1, Z_2? What if Z_1, Z_2 were also correlated?

To re-phrase: If $X_1, \dots, X_m$ can be linearly combined into $Y_1, \dots, Y_m$ where Y_i, Y_j are indpendent and "contain" all of the variation, then Y_i's are called the *principal components*.

(many other names exist too...)

We say $X_1, \dots, X_m$ have been *de-correlated*.

Consider $\vec{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_m \end{pmatrix}$, Y_i a random variable (stocks, radiation, EEG signals, etc.)

then the covariance matrix is $C_{ij} = \text{cov}(Y_i, Y_j)$

$$\Rightarrow \quad C = \mathbb{E}\left(\left(\vec{Y} - \mathbb{E}(\vec{Y})\right)\left(\vec{Y}^T - \mathbb{E}(\vec{Y}^T)\right) \right)$$

WLOG, assume that $\mathbb{E}(Y_i) = 0$ and $\text{Var}(Y_i) = 1$.

Recall: C is a SPsD (SPD) matrix, and therefore can be diagonalized:

$$C = WDW^T \quad \text{where} \quad W = (\vec{w}_1 \cdots \vec{w}_m) \quad \text{are}$$

the eigenvectors of C (with $\|\vec{w}_i\| = 1$) and

$D = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_m \end{pmatrix}$ are the eigenvalues.

The spectral representation of C is then:

$$C = \sum \lambda_j \vec{w}_j \vec{w}_j^T \qquad (\text{assume } \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 0).$$

Define: $Y_{(k)} = \vec{w}_k^T \vec{Y}$ ← projection of $\vec{Y}$ in the direction of maximum variance.

$$\Rightarrow \text{Var}(Y_{(k)}) = \text{Var}(\vec{w}_k^T \vec{Y})$$

$$= \text{Var}\left(\sum_{i=1}^{m} w_{ki} Y_i \right)$$

$$= \mathbb{E}\left(\sum_{i,j} w_{ki} Y_i Y_j w_{kj} \right)$$

$$= \vec{w}_k^T C \vec{w}_k = \vec{w}_k^T \lambda_k \vec{w}_k = \boxed{\lambda_k}$$

The "total variation" of $Y_1, \dots, Y_m$ can be defined as

$$\text{trace } C = \sum \lambda_i$$

And therefore $Y_{(k)}$ contains $\frac{\lambda_k}{\text{trace } C} = \frac{\lambda_k}{\sum \lambda_k}$ percent of the total variation.

To "approximate" the variety, pick p s.t. $\frac{\sum_1^p \lambda_k}{\sum_1^m \lambda_k}$ is large enough, and then form

$$\vec{Y}_p = \begin{pmatrix} Y_{(1)} \\ \vdots \\ Y_{(p)} \end{pmatrix} = \begin{pmatrix} w_1^T \\ \vdots \\ w_p^T \end{pmatrix} \begin{pmatrix} Y_1 \\ \vdots \\ Y_m \end{pmatrix}$$

← dimension reduction from m to p.

$\Rightarrow C_p = \mathbb{E}\left(\vec{Y}_p \vec{Y}_p^T\right)$ is the "approximate" covariance matrix.

Thm C_p is the matrix closest to C in the Frobenius norm.

This was for random variate $Y_1, \dots, Y_m$ — how about for actual data?

Eg. X_i models stocks, but you observe X_{ij}, $j = 1, \dots, n$ prices/returns, how do you construct an "index" which captures "most" of the variance of the market?

PCA from data

Let $X_{i1} \cdots X_{in}$, $i = 1, \ldots m$ be n realizations of each of the m random variables.

- Assume $\bar{X}_i = \frac{1}{n} \sum_{1}^{n} X_{ij} = 0$

- The sample covariance matrix is S, with

$$S_{ij} = \frac{1}{n-1} \sum X_{ik} X_{jk}$$

- Assume that $S_{ii} = 1$.

- Let X = "data matrix", $m \times n$, then $S = \frac{1}{n-1} X X^T$.

The matrix S is SPSD and therefore also diagonalizable,

$$S = W D W^T \quad \left(W, D \text{ are } \underline{\text{not}} \text{ identically the same as before} \right)$$

$$= \sum_{i=1}^{m} \lambda_i \vec{w}_i \vec{w}_i^T$$

And likewise the principal components of the data X are

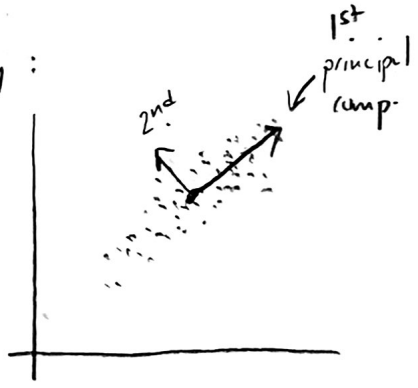
$$\tilde{X}_{(i)} = \vec{w}_i^T X$$

$1 \times n$ ← a row vector of "observations" of the principal component.

$1 \times m$ $m \times n$

w_{ij} = the contribution of

Graphically:



Practical Implementation of PCA

Recall, if the sample covariance matrix is

$$S = (X - \bar{X})(X - \bar{X})^T \frac{1}{n-1}$$

$$= W D W^T$$

However, we can work directly with the matrix $\frac{X - \bar{X}}{\sqrt{n-1}}$ by taking its SVD:

$$\frac{X - \bar{X}}{\sqrt{n-1}} = U S V^T$$

and then obviously:

$$\left(\frac{X - \bar{X}}{\sqrt{n-1}}\right)\left(\frac{X - \bar{X}}{\sqrt{n-1}}\right)^T = U S V^T V S^T U^T$$

$$= U S^2 U^T$$

$$\uparrow \quad \uparrow \quad \uparrow$$

$$W \quad D \quad W^T$$

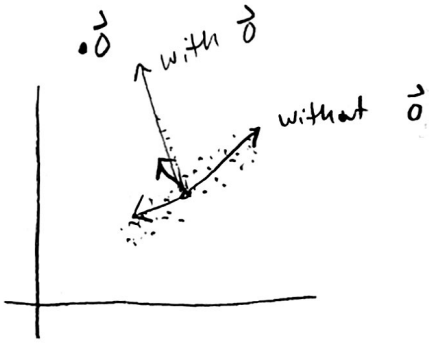
This is often much cheaper than forming S (mn^2) and then computing its eigenvectors (m^3), particularly if only one or two (or k) principal components are needed.

(Also, forming S squares the cond. number of $X - \bar{X}$, making it more accurately to work with the SVD instead.)

Robustness of PCA

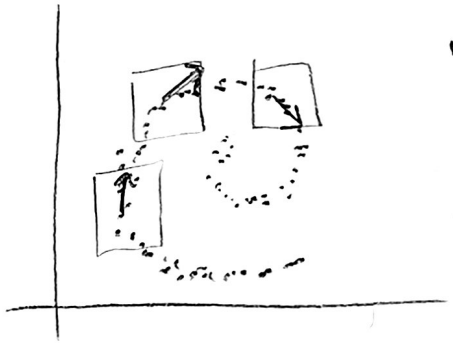
What types of things affect PCA?

Ex: outliers



Solution? Somehow ignore outliers:

1. remove them, and do global PCA
2. "locally cluster", and then do "local PCA"



What do PCs look like?

Clustering first avoids global effects.

Note: Much of this is "art"...

Related methods - Factor analysis (different generative model)

- Independent component analysis (different goal, independence) <+> principal