Principal Component Analysis

Karl Pearson (1857 - 1936)

MATH-412 - Statistical Machine Learning

Matrix multiplication as sums of column outer products

- Consider two matrices $A \in \mathbb{R}^{n \times K}$ and $B \in \mathbb{R}^{p \times K}$.
- Let $\boldsymbol{a}_k$ and $\boldsymbol{b}_k$ denote the kth column respectively of A and B, so that
- $A = \sum_{k=1}^{K} \boldsymbol{a}_k \boldsymbol{e}_k^\top$ and $B = \sum_{k=1}^{K} \boldsymbol{b}_k \boldsymbol{e}_k^\top$,
  where $\boldsymbol{e}_k \in \{0,1\}^K$ is the kth element of the canonical basis.

Lemma

$$AB^\top = \sum_{k=1}^{K} \boldsymbol{a}_k \boldsymbol{b}_k^\top \tag{†}$$

Proof: We have

$$AB^\top = \sum_{j=1}^{K} \boldsymbol{a}_j \boldsymbol{e}_j^\top \sum_{k=1}^{K} \boldsymbol{e}_k \boldsymbol{b}_k^\top = \sum_{j=1}^{K} \sum_{k=1}^{K} \boldsymbol{a}_j (\boldsymbol{e}_j^\top \boldsymbol{e}_k) \boldsymbol{b}_k^\top,$$

hence the result since $\boldsymbol{e}_j^\top \boldsymbol{e}_k = \delta_{j,k}$.

Empirical covariance and correlation

For centered vectors :

$$\widehat{\Sigma} = \frac{1}{n} X^\top X = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top$$

For non centered vectors :

$$\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top$$

Another common operation is to normalize the data by dividing each column of X by its standard deviation. This leads to the empirical correlation matrix.

$$C = \mathrm{Diag}(\widehat{\sigma})^{-1} \widehat{\Sigma}\, \mathrm{Diag}(\widehat{\sigma})^{-1} \qquad \text{with} \quad \widehat{\sigma}_k^2 = \widehat{\Sigma}_{k,k}.$$

$$C_{k,k'} = \frac{1}{n} \sum_{i=1}^n \Big(\frac{x_i^{(k)} - \bar{x}^{(k)}}{\widehat{\sigma}_k} \Big) \Big(\frac{x_i^{(k')} - \bar{x}^{(k')}}{\widehat{\sigma}_{k'}} \Big).$$

Normalisation is optional...

PCA from the analysis point of view

Data vectors live in $\mathbb{R}^p$ and one seeks a direction v in $\mathbb{R}^p$ such that the variance along this direction is maximal.

But, assuming centered data,

$$
\begin{aligned}
\mathrm{Var}((\boldsymbol{v}^\top \mathbf{x}_i)_{i=1\dots n}) &= \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{v}^\top \mathbf{x}_i)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{v}^\top \mathbf{x}_i\mathbf{x}_i^\top \boldsymbol{v} \\
&= \boldsymbol{v}^\top \Big(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top\Big)\boldsymbol{v} \\
&= \boldsymbol{v}^\top \widehat{\Sigma}\, \boldsymbol{v}
\end{aligned}
$$

One needs to solve

$$\max_{\|\boldsymbol{v}\|_2=1} \boldsymbol{v}^\top \widehat{\Sigma}\, \boldsymbol{v}$$

Solution:

- First eigenvectors of $\widehat{\Sigma}$.
- Let's call it $\boldsymbol{v}_1$.

Deflation

What is the second best direction to project the data on in order to maximize the variance ?

One can perform a deflation

$$\forall i, \quad \widetilde{\mathbf{x}}_i \leftarrow \mathbf{x}_i - \boldsymbol{v}_1(\boldsymbol{v}_1^\top \mathbf{x}_i)$$

Which translates at the matrix level by: $\widetilde{X} \leftarrow X - X\boldsymbol{v}_1\boldsymbol{v}_1^\top.$

Then again find the direction of maximal variance. So with

$$\widetilde{\widehat{\Sigma}} = \frac{1}{n}\widetilde{X}^\top \widetilde{X},$$

we solve $\max_{\|\boldsymbol{v}\|_2} \boldsymbol{v}^\top \widetilde{\widehat{\Sigma}}\, \boldsymbol{v}$

Or equivalently $\max_{\|\boldsymbol{v}\|_2} \boldsymbol{v}^\top \widehat{\Sigma}\, \boldsymbol{v} \quad \text{s.t.} \quad \boldsymbol{v} \perp \boldsymbol{v}_1.$

Solution: This yields the second eigenvector of $\widehat{\Sigma}$, say $\boldsymbol{v}_2$. Etc.

Principal directions

We usually call

- **principal directions (or factors)** of the points cloud the vectors

$$\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_k.$$

- k**the principal component (or scores)**:
  the projection of the data on the k principal direction.

$$(\boldsymbol{v}_k^\top \mathbf{x}_i)_{i=1\ldots n}$$

The principal directions are the eigenvectors of $\widehat{\Sigma} = \tilde{V}\, S_E^2 \tilde{V}^\top$.

Singular value decomposition (SVD)

- Principal directions also appear in singular value decomposition of data matrix itself: $X = \tilde{U}\tilde{S}\tilde{V}^\top$, with
- $\tilde{U} \in \mathbb{R}^{n\times n}$ orthogonal in $\mathbb{R}^n$
- $\tilde{S} \in \mathbb{R}^{n\times p}$ a (rectangular) diagonal matrix .
- $\tilde{V} \in \mathbb{R}^{p\times p}$ orthogonal in $\mathbb{R}^p$

Reduced SVD

Often more convenient to look at $X = USV^\top$ with,

- $U \in \mathbb{R}^{n\times r}$ whose columns are orthonormal.
- $S \in \mathbb{R}^{r\times r}$ squared diagonal strictly positive.
- $V \in \mathbb{R}^{p\times r}$ whose columns are orthonormal.
- r is the rank of X

If the diagonal of S is such that $s_1 > s_2 > \ldots > s_r > 0$, then the reduced SVD is unique up to column signs of U. $S_E \in \mathbb{R}^{p\times p}$ completes S by adding zeroes.

Simultaneous optimisation

Let $X = USV^\top$ be the (reduced) SVD of X, and

- $U_{[k]} \in \mathbb{R}^{n \times k}$ the matrix formed by the first k columns of U
- $V_{[k]} \in \mathbb{R}^{p \times k}$ the matrix formed by the first k columns of V
- $S_{[k]} \in \mathbb{R}^{k \times k}$ the diagonal matrix with the first (largest) k singular values in S

Theorem (Eckart-Young)

The solution of

$$\min_Z \|X - Z\|_F^2 \quad \textit{s.t.} \quad \operatorname{rank}(Z) \leq k$$

is

$$Z = X_{[k]} \qquad \textit{with} \quad X_{[k]} := U_{[k]} S_{[k]} V_{[k]}^\top .$$

Can be interpreted as projection of X on columns of $V_{[k]}$

Orthogonal projection on the principal subspace

Let

- $V = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k] \in \mathbb{R}^{p\times k}$ be a matrix of orthonormal columns,
- $\mathcal{V}_k = \mathrm{span}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k) \subseteq \mathbb{R}^p$,
- $\mathrm{Proj}_{\mathcal{V}_k}(\mathbf{x})$ be the projection of $\mathbf{x} \in \mathbb{R}^p$ on $\mathcal{V}_k$,

then

$$\mathrm{Proj}_{\mathcal{V}_k}(\mathbf{x}) = VV^\top \mathbf{x} \overset{(\dagger)}{=} \sum_{j=1}^{k} \boldsymbol{v}_j \boldsymbol{v}_j^\top \mathbf{x}.$$

Interpretation:

- The sum of the projections on the $\boldsymbol{v}_k$s is equal to the projection on $\mathcal{V}_k$.
- This is of course the main property that we seek in an orthonormal basis.

The design matrix with the projections of all the dataset is therefore $XVV^\top$.

SVD factorization via outer products

Given that S is a diagonal matrix, we have $US = [s_1\boldsymbol{u}_1, s_2\boldsymbol{u}_2, \ldots, s_r\boldsymbol{u}_r]$.
So by (†)

$$X = USV^\top = (US)V^\top = \sum_{j=1}^{r} s_j \boldsymbol{u}_j \boldsymbol{v}_j^\top.$$

The projection of the data on the space spanned by the k first principal directions is

$$XV_{[k]}V_{[k]}^\top = \sum_{j=1}^{r} s_j \boldsymbol{u}_j \boldsymbol{v}_j^\top V_{[k]}V_{[k]}^\top = U_{[k]}S_{[k]}V_{[k]}^\top V_{[k]}V_{[k]}^\top = U_{[k]}S_{[k]}V_{[k]}^\top = \sum_{j=1}^{k} s_j \boldsymbol{u}_j \boldsymbol{v}_j^\top.$$

The matrix of the first k principal components is thus $XV_{[k]} = USV^\top V_{[k]} = U_{[k]}S_{[k]}$.
The kth principal component (score) of $\mathbf{x}_i$ is $\mathbf{x}_i^\top \boldsymbol{v} = s_k\, u_i^{(k)}$

Two different views of PCA

Given data matrix $X = (x_1^\top, \ldots, x_n^\top)^\top \in \mathbb{R}^{n\times p}$,

Analysis view

Find projection $v \in \mathbb{R}^p$ maximizing variance:

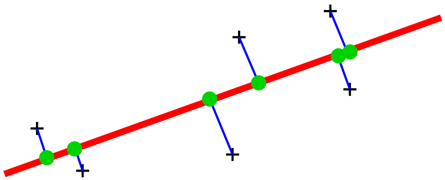
$$\begin{array}{ll} \max_{v\in\mathbb{R}^p} & v^\top X^\top X\, v \\ \text{s.t.} & \|v\|_2 \leq 1 \end{array}$$

$\rightarrow$ deflate and iterate to obtain more components.

Synthesis view

Find $V = [v_1, \ldots, v_k]$ s.t. x_i have low reconstruction error on span(V):

$$\min_{b_i, v_i \in \mathbb{R}^p} \left\| X - \sum_{i=1}^{k} b_i v_i^\top \right\|_F^2$$



Interpretation

- PCA basically represents a change-of-basis
- In the new basis, everything is mathematically simpler
- But our intuition/interest is in terms of original basis
- Coordinates in original basis correspond to variables/features (age, weight, height, ...)
- Coordinates in PCA basis are linear combinations of variables/features: (e.g., 0.3*age + 0.6*weight + 0.89*height
- Can have sparse combinations by penalisation

$$\arg\max_{\|\mathrm{v}\|=1} \mathrm{v}^t \widehat{\Sigma} \mathrm{v} + \lambda \|\mathrm{v}\|_1$$

- PCA depends on scale (height in cm / m changes everything)
- If units are very different can normalise and work with correlation matrix
- Otherwise can have expert knowledge

Number of components

- A priori, there is no unequivocal way to choose a truncation level k
- Often use % of variance explained:
- The variance of i-th coordinate is $\widehat{\Sigma}_{ii}$
- total variance is tr$\widehat{\Sigma} = \sum s_{ii}^2$
- Look at $\sum_{i=1}^k s_{ii}^2$ and stop when it is $\geq \beta$tr$\widehat{\Sigma}$, e.g., $\beta = 85\%$
- Or plot s_{ii}^2 and look for an "elbow"