Principal Component Analysis (PCA)

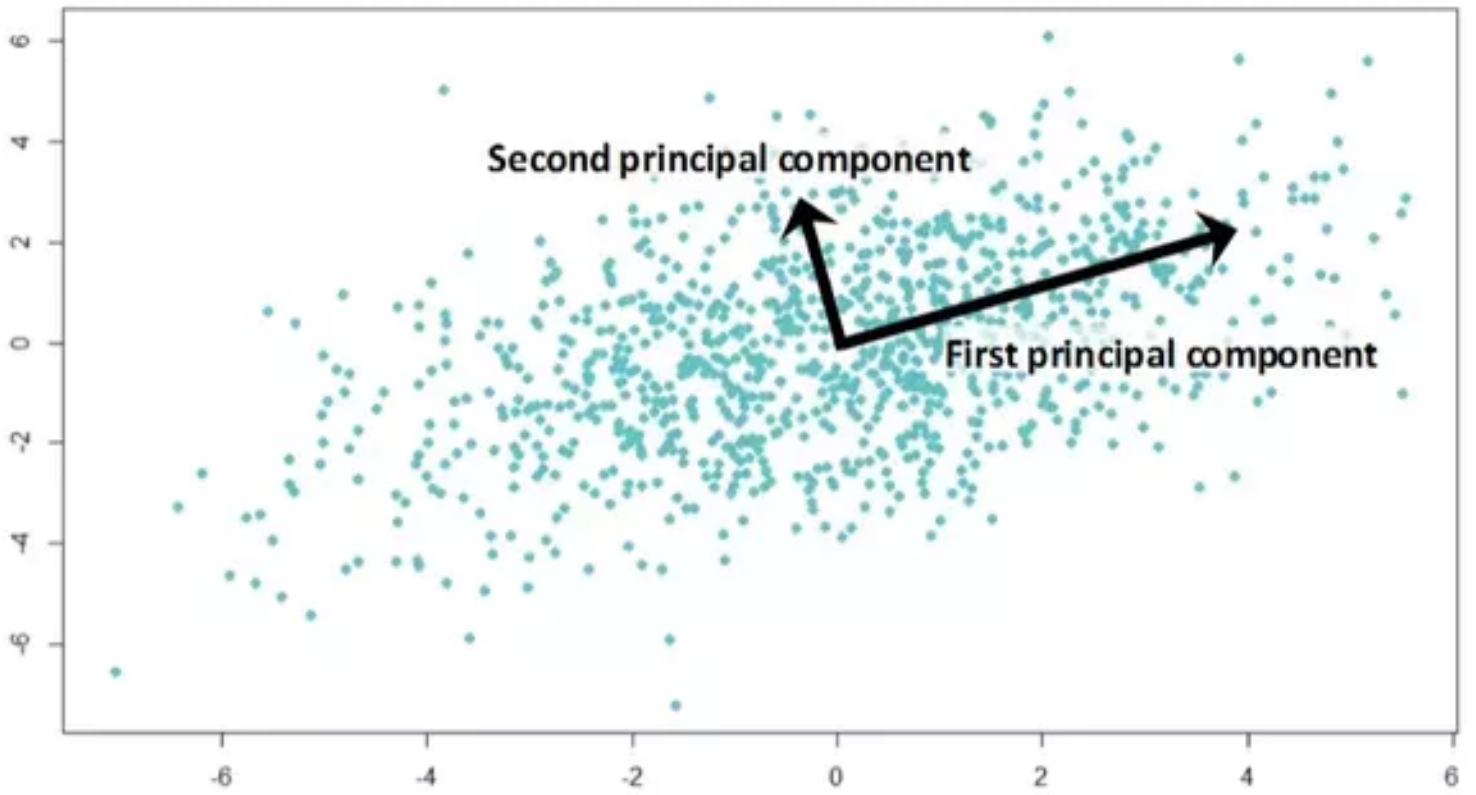
- Imagine we want to predict the results of the next election. There are many variables to base our prediction on, eg. the recent trend in housing prices, the local and global stock markets, crime rate, wages, recent changes in taxes etc.
- From our lesson on regression, we know we can fit all this data into a design matrix X, where the data-points $x^{(i)}$ are the rows of X, each of which are vectors in $\mathbb{R}^p$ (thus X has p columns).

- Each component in $x^{(i)}$, and there are p of them, represents a particular variable you want to model.
- What if p is too large and we want to only consider the most important L ($< p$) variables?
- What if the variables are not independent from one another but are correlated in a way that is unknown to us?

- Rather than simply eliminating some of the variables haphazardly, we want to perform a transformation on the data-points such that after the transformation, the new variables/components are
  - (i) independent from one another
  - (ii) ranked in terms of their importance (axes with the most variation), so we can choose the L most important ones (dimensionality reduction)

- Note that one potential drawback is that such a transformation would lead to a loss in interpretability of the data.

Principal components



Singular value decomposition (SVD)

Theorem

Given any real n x p matrix M, there exists a factorization of M of the form

$$M = U\Sigma V^T$$

where

- *U is an orthogonal n x n matrix,*
- *Σ is a n x p matrix with non-negative real numbers on the diagonal and zero elsewhere,*
- *V is an orthogonal p x p matrix.*

The diagonal entries of Σ are called the singular values of M.

PCA steps

Given a design matrix X, perform the following steps:

(1) For each column of X, compute the mean and subtract it from each entry in the column. At the same time, compute the standard deviation and divide each entry in the column by it (this ensures that each column is normalized to have mean 0 and standard deviation 1).

(2) Do SVD on X to yield $X = U\Sigma V^T$, then multiply X by V to get $XV = U\Sigma$.

(3) Retain the first L columns of XV and remove the remaining columns.

Why does PCA work?

We want to find a transformation $x^{(i)}Z = t^{(i)} = \left[t_1^{(i)}, \ldots, t_p^{(i)}\right]$ where the variance of the first component

$$t_1^{(i)} = \left\langle x^{(i)}, z_1 \right\rangle \quad (z_1 \text{ is the first column of Z})$$

is maximized, i.e.

$$z_1 = \arg\max_{\|z\|=1} \sum_{i=1}^{n} \left(t_1^{(i)}\right)^2 = \arg\max_{\|z\|=1} \sum_{i=1}^{n} \left(\left\langle x^{(i)}, z \right\rangle\right)^2$$
$$= \arg\max_{\|z\|=1} \|Xz\|^2 = \arg\max_{\|z\|=1} z^T X^T X z.$$

- Let $\lambda_1, \ldots, \lambda_p$ denote the eigenvalues of $X^T X$ arranged in descending order, and let $v_1, \ldots, v_p$ denote the corresponding eigenvectors. As the eigenvectors form an orthonormal basis, we can represent

$$z = a_1 v_1 + \ldots + a_p v_p,$$

  where $a_i = \langle z, v_i \rangle$ for all i.
- Then we have

$$\begin{aligned} z^T X^T X z &= \left\langle z, X^T X z \right\rangle \\ &= \langle a_1 v_1 + \ldots + a_p v_p,\ a_1 \lambda_1 v_1 + \ldots + a_p \lambda_p v_p \rangle \\ &= a_1^2 \lambda_1 + \ldots + a_p^2 \lambda_p, \end{aligned}$$

  and essentially we need to find $(a_1, \ldots, a_p)$ with $\sum_i a_i^2 = 1$ which maximizes this expression.

- Hence, the vector which maximizes the variance of the first component must be the eigenvector v_1, where $(a_1, \ldots, a_p) = (1, \ldots, 0)$, of $X^T X$ corresponding to its largest eigenvalue λ_1.
- Similarly, to compute the next most variable component $t_2^{(i)}$ one must compute the second eigenvector of $X^T X$, and so on.
- $X^T X$ is proportional to the sample covariance matrix of the original dataset, and recall that we need to first compute it to obtain V in the SVD of X.
- Thus, we see that V is precisely the transformation Z we are looking for.