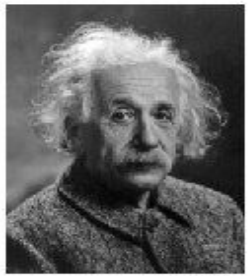
Redução de Dimensionalidade

PCA

What is Principal Component Analysis? What does it do?

So, first let's build some intuition.

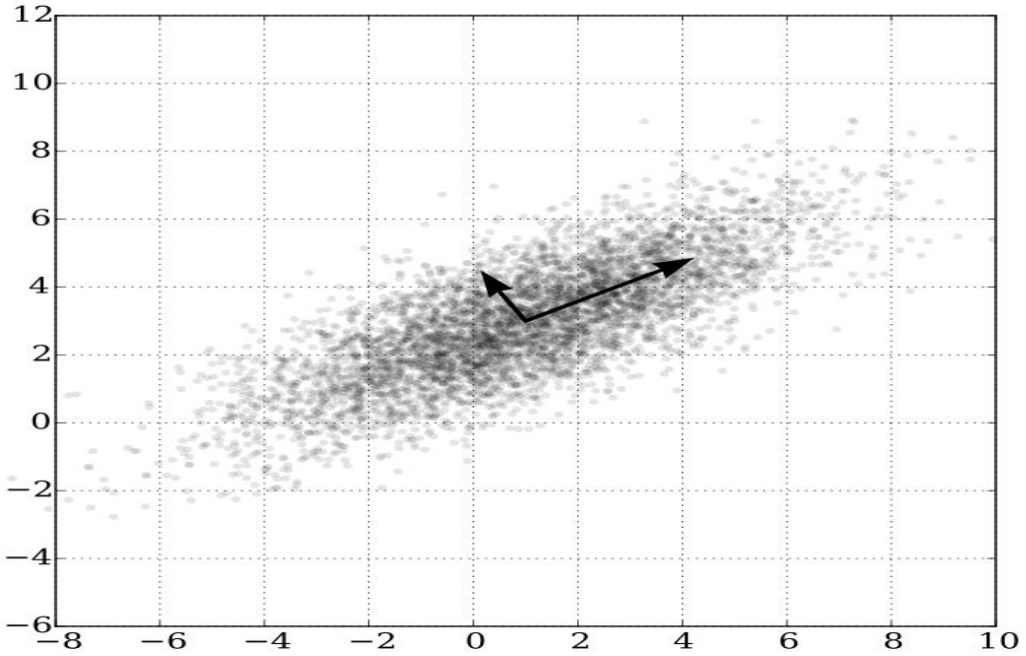
"You do not really understand something unless you can explain it to your grandmother", Albert Einstein



wikipedia

With thanks for many of these explanations to http://stats.stackexchange.com/questions/2691/making-sense-of-principal-component-analysis-eigenvectors-eigenvalues

Principal Component Analysis



PCA

- Principal components analysis (PCA) is a technique that can be used to simplify a dataset
- It is a linear transformation that chooses a new coordinate system for the data set such that greatest variance by any projection of the data set comes to lie on the first axis (then called the first principal component), the second greatest variance on the second axis, and so on.
- PCA can be used for reducing dimensionality by eliminating the later principal components.

PCA

- By finding the eigenvalues and eigenvectors of the covariance matrix, we find that the eigenvectors with the largest eigenvalues correspond to the dimensions that have the strongest correlation in the dataset.
- This is the principal component.
- PCA is a useful statistical technique that has found application in: – fields such as face recognition and image compression – finding patterns in data of high dimension.

$$\frac{1}{n}\sum_{t=1}^{n}\|\hat{\mathbf{x}}_t - \mathbf{x}_t\|_2^2 = \frac{1}{n}\sum_{t=1}^{n}\sum_{j=k+1}^{d}\mathbf{w}_j^\top(\mathbf{x}_t-\mu)(\mathbf{x}_t-\mu)^\top\mathbf{w}_j = \sum_{j=k+1}^{d}\mathbf{w}_j^\top\Sigma\mathbf{w}_j$$

$$\operatorname*{argmin}_{\mathbf{w}:\|\mathbf{w}_j\|_2=1}\sum_{j=k+1}^{d}\mathbf{w}_j^\top\Sigma\mathbf{w}_j = \sum_{j=k+1}^{d}\lambda_j\mathbf{w}_j^\top\mathbf{w}_j = \sum_{j=k+1}^{d}\lambda_j$$

Solution : $\mathbf{w}_j$'s are eigenvectors and λ_j's are corresponding eigenvalues

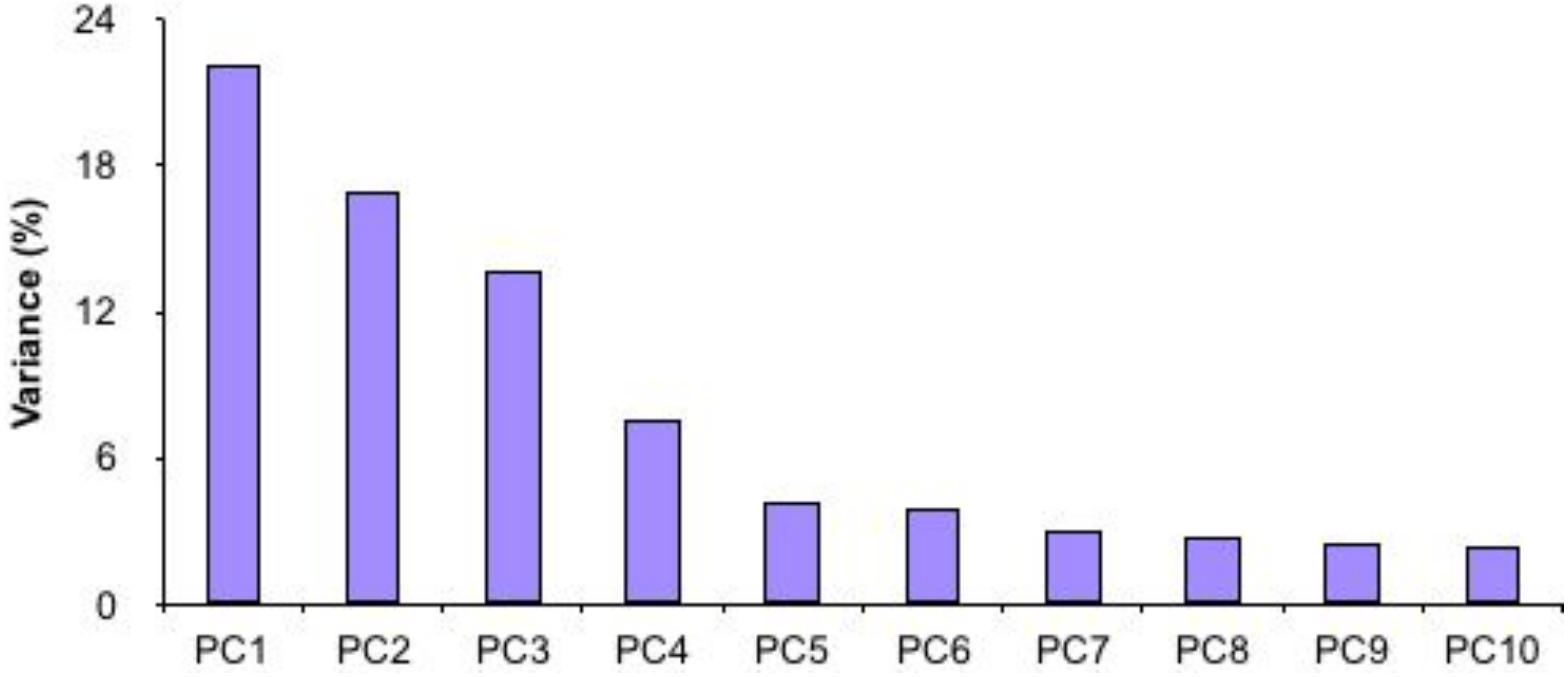
Clearly to minimize reconstruction error, we need to minimize $\sum_{j=k+1}^{d}\lambda_j$. In other words we discard the $d-k$ directions that have the smallest eigenvalue

$$\Sigma[i,j] = \frac{1}{n}\begin{bmatrix} \mathbf{x}_1[i] - \mu[i] \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{x}_n[i] - \mu[i] \end{bmatrix}^{\top}\begin{bmatrix} \mathbf{x}_1[j] - \mu[j] \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{x}_n[j] - \mu[j] \end{bmatrix}$$

$$\operatorname{cov}(A,B) = \mathbb{E}[(A - \mathbb{E}[A])(B - \mathbb{E}[B])]$$

$$W = \operatorname{SVD}(X - \mu, K)$$

PCA



PCA

Average squared projection error :

Total Variance in data :

If those options don't seem equivalent to you…