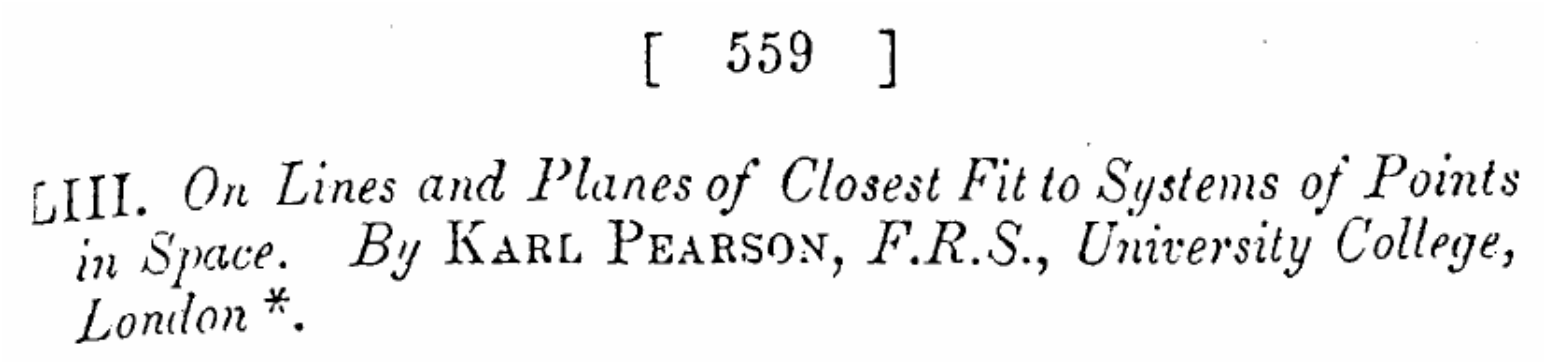
Principal Component Analysis – the original paper

ADRIAN FLOREA

PAPERS WE LOVE – BUCHAREST CHAPTER – MEETUP #12

DECEMBER 16TH, 2016

Karl Pearson, 1901



[559]

LIII. *On Lines and Planes of Closest Fit to Systems of Points in Space.* By KARL PEARSON, *F.R.S., University College, London* *.

K. Pearson, "*On lines and planes of closest fit to systems of points in space*",
The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science, Sixth Series, 2, pp. 559-572 (1901)

(1857-1936)

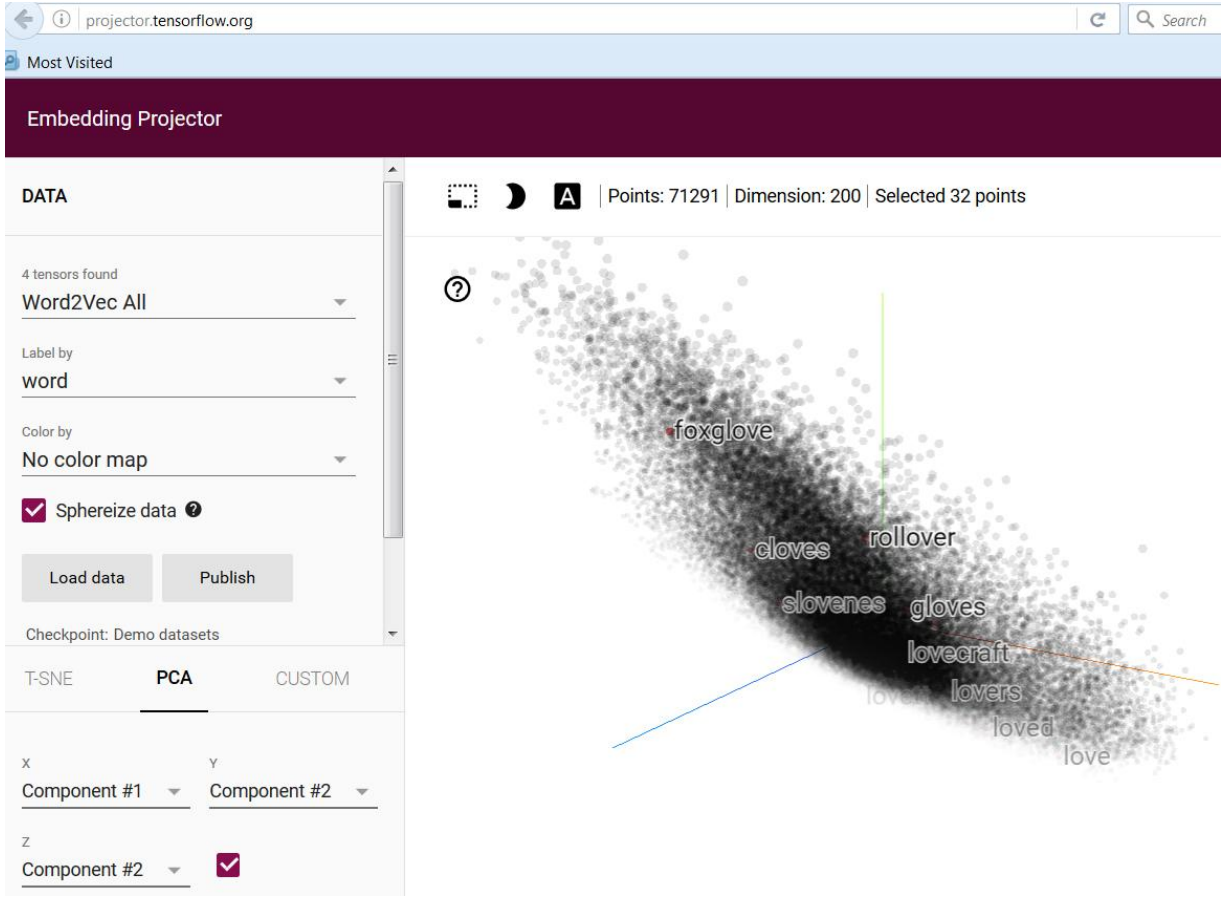
Dimensionality reduction (Pearson's words)

(1) IN many physical, statistical, and biological investigations it is desirable to represent a system of points in plane, three, or higher dimensioned space by the "best-fitting" straight line or plane. Analytically this

(1901)

These two illustrations may suffice to show that the methods of this paper can be easily applied to numerical problems; the labour is not largely increased if we have a considerable number of points. It becomes more cumbersome if we have four, five, or more variables or characters which involve the

Dimensionality reduction (115 years later)

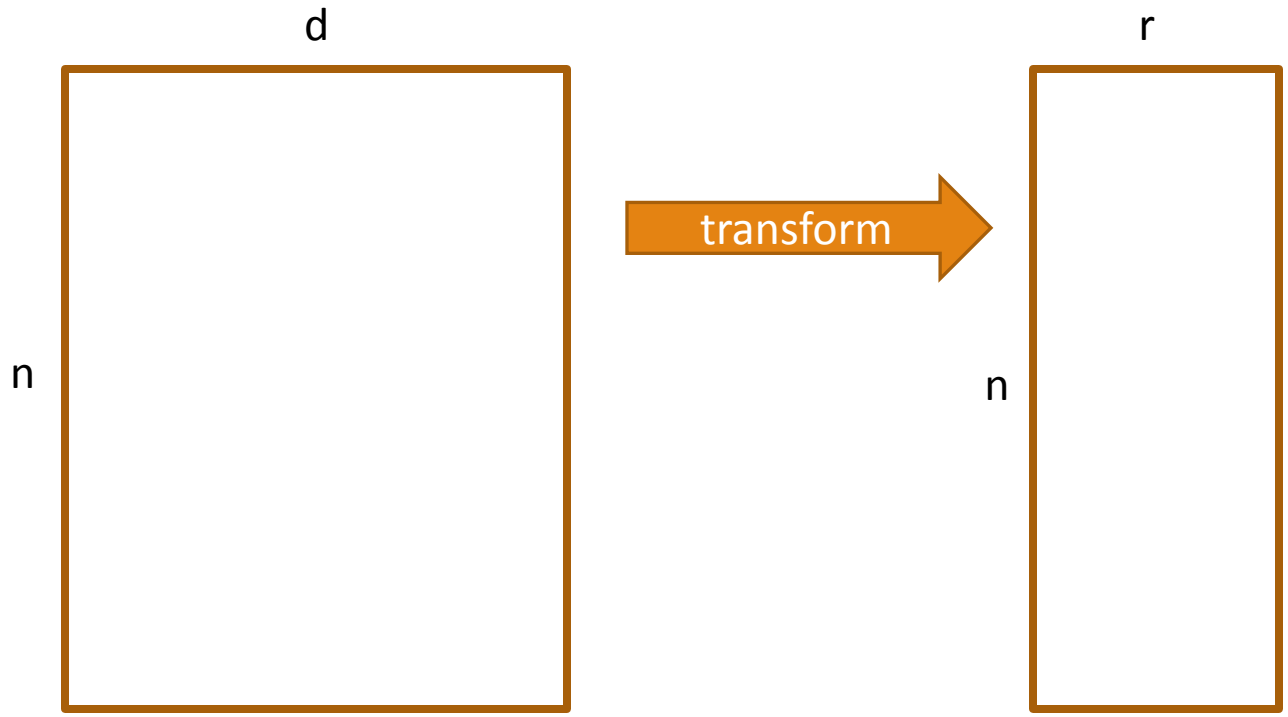


(2016)

71291 points
200 features
2 principal components

https://projector.tensorflow.org

Dimensionality reduction



Original data set

- n observations
- d possibly correlated features

New data set

- r << d
- n observations
- r uncorrelated features
- retain most of the data variability

Introduction

$$\overline{\overline{X}} = (\overline{x_1}\,\overline{x_2} \ldots \overline{x_n})^T$$

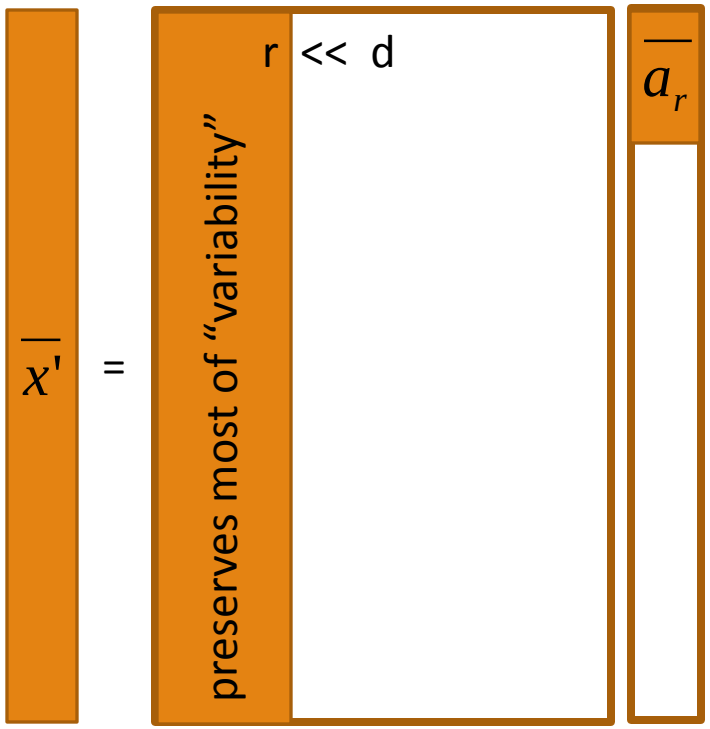
$$\overline{x_i} = (x_{i1} x_{i2} \ldots x_{id})^T \in R^d, i = \overline{1,n}$$

$$\overline{x} \in R^d \Rightarrow \overline{x} \in \text{span}(\overline{e_1}, \overline{e_2}, \ldots, \overline{e_d})$$

$$\exists \{\overline{u_i} \in R^d \mid i = \overline{1,d}\}, \overline{u_i}^T \overline{u_j} = \delta_{i,j}, \text{s.t.} \overline{x} \in \text{span}(\overline{u_1}, \overline{u_2}, \ldots, \overline{u_d})$$

$$\left.\begin{array}{l} \overline{x} \in \text{span}(\overline{u_1}, \overline{u_2}, \ldots, \overline{u_d}) \Rightarrow \overline{x} = a_1 \overline{u_1} + a_2 \overline{u_2} + \ldots + a_d \overline{u_d} \Rightarrow \overline{x} = \overline{\overline{U}}\,\overline{a} \\ \text{Columns of } \overline{\overline{U}} \text{ are an orthonormal basis} \Rightarrow \overline{\overline{U}} \text{ is orthogonal} => \overline{\overline{U}}^{-1} = \overline{\overline{U}}^T \end{array}\right\} \Rightarrow \overline{\overline{U}}^T \overline{x} = \overline{\overline{U}}^T \overline{\overline{U}}\,\overline{a} \Rightarrow \overline{a} = \overline{\overline{U}}^T \overline{x}$$

Goal: find an “optimal” basis



$$\frac{(x_1 x_2 \ldots x_d)^T}{span(\overline{e_1}, \overline{e_2}, \ldots, \overline{e_d})} \xrightarrow{\overline{\overline{U}}^T} \frac{(a_1 a_2 \ldots a_d)^T}{span(\overline{u_1}, \overline{u_2}, \ldots, \overline{u_d})}$$

$$\left.\begin{aligned} &\overline{x'} = a_1\overline{u_1} + a_2\overline{u_2} + \ldots + a_r\overline{u_r} \\ &\overline{x'} = \overline{\overline{U_r}}\,\overline{a_r} \\ &\overline{a} = \overline{\overline{U}}^T\,\overline{x} \Rightarrow \overline{a_r} = \overline{\overline{U_r}}^T\,\overline{x} \end{aligned}\right\} \Rightarrow \overline{x'} = \overline{\overline{U_r}}\,\overline{\overline{U_r}}^T\,\overline{x} \Rightarrow \overline{x'} = \overline{\overline{P_r}}\,\overline{x}$$

r=1

Find unitary $\bar{u}, \bar{u}^T \bar{u} = 1$ that maximizes the variance of the projected points on it. Assume $\mu_{\bar{u}} = 0$

Projection of $\bar{x}_i$ on $\bar{u}$:

$$\| \bar{x'}_i \| = \| \bar{x}_i \| \cos(\bar{x}, \bar{u}) = \| \bar{x}_i \| \frac{\bar{u}^T \bar{x}_i}{\| \bar{u} \| \cdot \| \bar{x}_i \|} = \bar{u}^T \bar{x}_i \Rightarrow \bar{x'}_i = (\bar{u}^T \bar{x}_i)\bar{u} = a_i \bar{u}$$

$$\sigma_{\bar{u}}^2 = \frac{1}{n} \sum_{i=1}^{n} (a_i - \mu_{\bar{u}})^2 = \frac{1}{n} \sum_{i=1}^{n} (\bar{u}^T \bar{x}_i - 0)^2 = \frac{1}{n} \sum_{i=1}^{n} (\bar{u}^T \bar{x}_i)^2 = \frac{1}{n} \sum_{i=1}^{n} (\bar{u}^T \bar{x}_i)(\bar{u}^T \bar{x}_i)^T$$

$$\sigma_{\bar{u}}^2 = \frac{1}{n} \sum_{i=1}^{n} \bar{u}^T (\bar{x}_i \bar{x}_i^T) \bar{u} = \bar{u}^T \left(\frac{1}{n} \sum_{i=1}^{n} \bar{x}_i \bar{x}_i^T \right) \bar{u} = \bar{u}^T \operatorname{cov}(\bar{\bar{X}}) \bar{u} = \bar{u}^T \bar{\bar{\Sigma}} \bar{u}$$

$$\max_{\bar{u}} \bar{u}^T \bar{\bar{\Sigma}} \bar{u}, \ \bar{u}^T \bar{u} = 1$$

First principal component

$$L(\bar{u},\alpha) = \bar{u}^T \bar{\bar{\Sigma}} \bar{u} - \alpha(\bar{u}^T \bar{u} - 1), \max_{\bar{u},\alpha} L(\bar{u},\alpha)$$

$$\frac{\partial L(\bar{u},\alpha)}{\partial \bar{u}} = \bar{0} \Rightarrow \frac{\partial}{\partial \bar{u}}(\bar{u}^T \bar{\bar{\Sigma}} \bar{u} - \alpha \bar{u}^T \bar{u} + \alpha) = \bar{0} \Rightarrow 2\bar{\bar{\Sigma}} \bar{u} - 2\alpha \bar{u} = \bar{0}$$

$\bar{\bar{\Sigma}} \bar{u} = \alpha \bar{u}$ α is an eigenvalue of the covariance matrix with the associated eigenvector $\bar{u}$

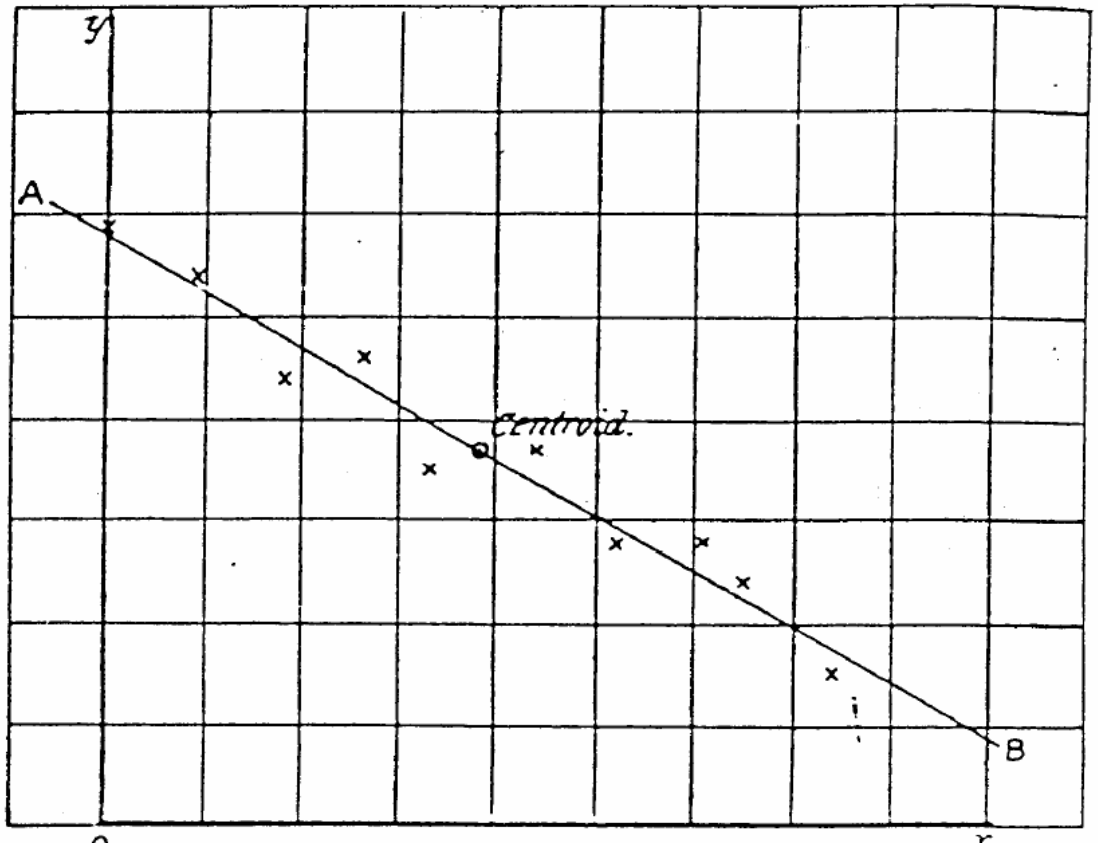
$$\sigma_{\bar{u}}^2 = \bar{u}^T \bar{\bar{\Sigma}} \bar{u} = \bar{u}^T \alpha \bar{u} = \alpha \Rightarrow \max_{\bar{u}} \sigma_{\bar{u}}^2 = \max_{\bar{u}} \alpha = \lambda_1$$

The largest eigenvalue is the projected variance on the associated eigenvector that is the direction of most variance, called the **first principal component**.

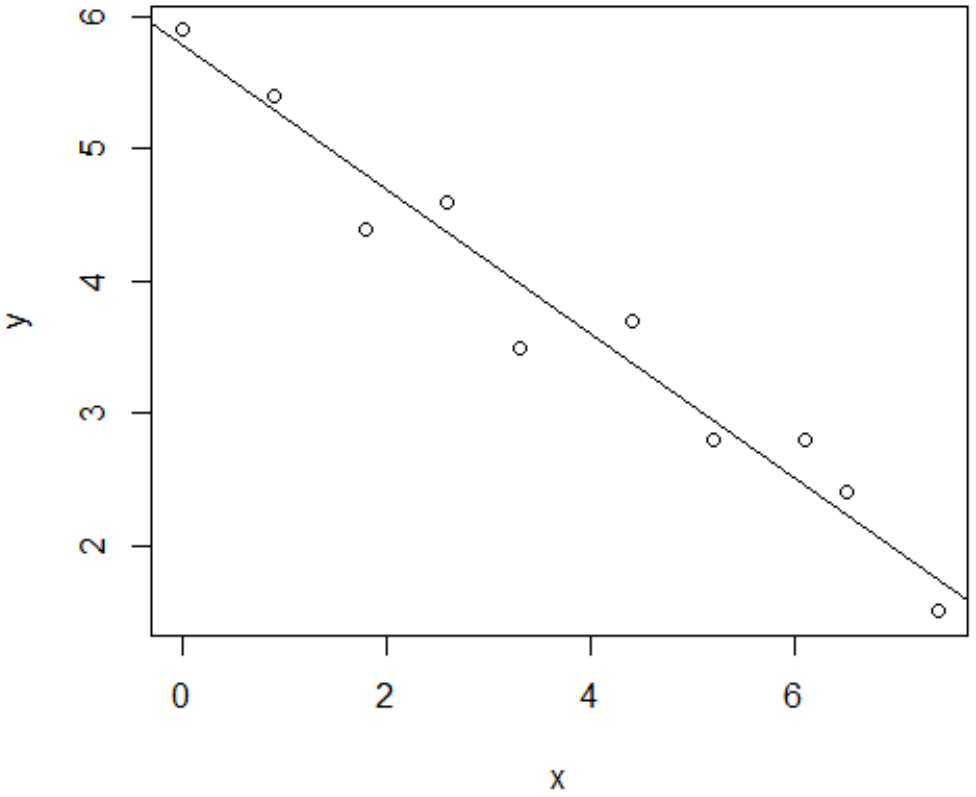
(7) *Numerical Illustrations.*

Case (i.). Find the best fitting straight line to the following system of points supposed of equal weight :

$x = 0$	$y = 5{\cdot}9$	$x = 4{\cdot}4$	$y = 3{\cdot}7$
$x = {\cdot}9$	$y = 5{\cdot}4$	$x = 5{\cdot}2$	$y = 2{\cdot}8$
$x = 1{\cdot}8$	$y = 4{\cdot}4$	$x = 6{\cdot}1$	$y = 2{\cdot}8$
$x = 2{\cdot}6$	$y = 4{\cdot}6$	$x = 6{\cdot}5$	$y = 2{\cdot}4$
$x = 3{\cdot}3$	$y = 3{\cdot}5$	$x = 7{\cdot}4$	$y = 1{\cdot}5$




```
d <- data.frame(x=c(0, 0.9, 1.8, 2.6, 3.3, 4.4, 5.2, 6.1, 6.5, 7.4),
                y=c(5.9, 5.4, 4.4, 4.6, 3.5, 3.7, 2.8, 2.8, 2.4, 1.5))
s <- cov(d)
e <- eigen(s)
pca <- princomp(d)
plot(d)
```



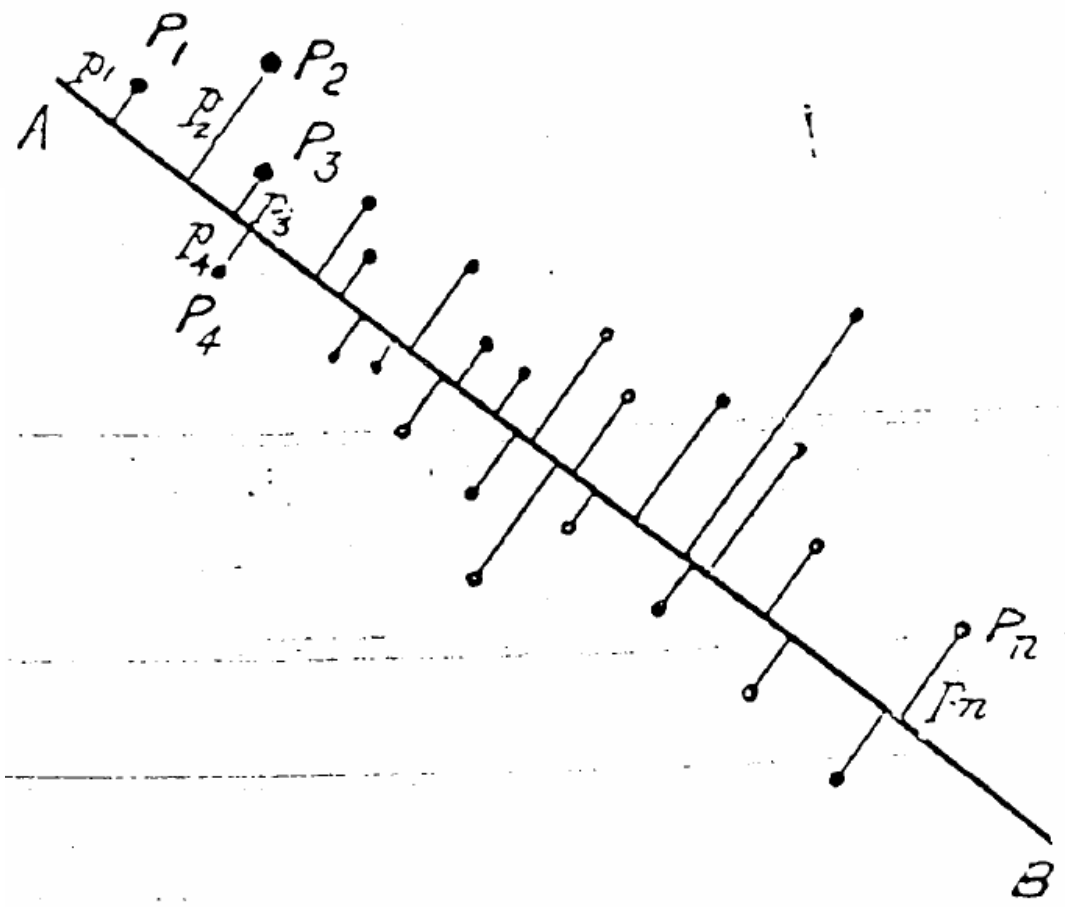
Of course the term "best fit" is really arbitrary; but a good fit will clearly be obtained if we make the sum of the squares of the perpendiculars from the system of points upon the line or plane a minimum.

For example:—Let $P_1, P_2, \ldots P_n$ be the system of points with coordinates $x_1, y_1; x_2, y_2; \ldots x_n y_n$, and perpendicular distances $p_1, p_2, \ldots p_n$ from a line A B. Then we shall make

$$U = S(p^2) = \text{a minimum.}$$

$$\begin{cases} \min\limits_{\bar{u}} MSE(\bar{u}) \\ \bar{u}^T \bar{u} = 1 \end{cases}$$

The equivalent optimization problem solved by Pearson



An equivalent objective function

$$MSE(\bar{u}) = \frac{1}{n}\sum_{i=1}^{n} \| \bar{x}_i - \bar{x'}_i \|^2 = \frac{1}{n}\sum_{i=1}^{n} (\bar{x}_i - \bar{x'}_i)^T (\bar{x}_i - \bar{x'}_i) = \frac{1}{n}\sum_{i=1}^{n} (\| \bar{x}_i \|^2 - 2\bar{x}_i^T \bar{x'}_i + \bar{x'}_i^T \bar{x'}_i)$$

But: $\bar{x'}_i = (\bar{u}^T \bar{x}_i)\bar{u} \Rightarrow MSE(\bar{u}) = \frac{1}{n}\sum_{i=1}^{n} (\| x_i \|^2 - 2\bar{x}_i^T (\bar{u}^T \bar{x}_i)\bar{u} + ((\bar{u}^T \bar{x}_i)\bar{u})^T (\bar{u}^T \bar{x}_i)\bar{u})$

$$= \frac{1}{n}\sum_{i=1}^{n} (\| x_i \|^2 - 2(\bar{u}^T \bar{x}_i)(\bar{x}_i^T \bar{u}) + (\bar{u}^T \bar{x}_i)(\bar{x}_i^T \bar{u})\bar{u}^T \bar{u}) = \frac{1}{n}\sum_{i=1}^{n} (\| x_i \|^2 - (\bar{u}^T \bar{x}_i)(\bar{x}_i^T \bar{u}))$$

$$= \frac{1}{n}\sum_{i=1}^{n} (\| x_i \|^2 - \bar{u}^T (\bar{x}_i \bar{x}_i^T)\bar{u}) = -\bar{u}^T \bar{\bar{\Sigma}} \bar{u} + \underbrace{\sum_{i=1}^{n} \frac{\| x_i \|^2}{n}}_{\text{const}} \Rightarrow \min_{\bar{u}} MSE(\bar{u}) = \min_{\bar{u}} (-\bar{u}^T \bar{\bar{\Sigma}} \bar{u}) = \max_{\bar{u}} \bar{u}^T \bar{\bar{\Sigma}} \bar{u} = \max_{\bar{u}} \sigma_{\bar{u}}^2$$

$$\min_{\bar{u}} MSE(\bar{u}) = \max_{\bar{u}} \sigma_{\bar{u}}^2$$

Maximizing the projected variance is **equivalent** with minimizing the mean squared error

r=2

We have found the vector $\bar{u}_1$ of the most variance (r = 1 case)

$$\begin{cases} \max_{\bar{v}} \sigma_{\bar{v}}^2 \\ \bar{v}^T \bar{v} = 1 \\ \bar{v}^T \bar{u}_1 = 0 \end{cases}$$

$$L(\bar{v}, \alpha, \beta) = \bar{v}^T \bar{\bar{\Sigma}} \bar{v} - \alpha(\bar{v}^T \bar{v} - 1) - \beta \bar{v}^T \bar{u}_1$$

$$\frac{\partial L}{\partial \bar{v}} = \bar{0} \Rightarrow 2\bar{\bar{\Sigma}}\bar{v} - 2\alpha\bar{v} - \beta\bar{u}_1 = \bar{0} \Rightarrow 2\bar{u}_1^T \bar{\bar{\Sigma}} \bar{v} = 2\alpha \bar{u}_1^T \bar{v} + \beta \bar{u}_1^T \bar{u}_1 \Rightarrow$$

$$\beta = 2\bar{u}_1^T \bar{\bar{\Sigma}} \bar{v} = 2\bar{v}\bar{\bar{\Sigma}}\bar{u}_1^T = 2\bar{v}^T \lambda_1 \bar{u}_1 = 2\lambda_1 \bar{v}^T \bar{u}_1 = 0 \Rightarrow 2\bar{u}_1^T \bar{\bar{\Sigma}} \bar{v} = 2\alpha \bar{u}_1^T \bar{v} \Rightarrow$$

$\bar{\bar{\Sigma}}\bar{v} = \alpha\bar{v}$ So $\bar{v}$ is an eigenvector of $\bar{\bar{\Sigma}}$

$\max_{\bar{v}} \sigma_{\bar{v}}^2 = \max_{\bar{u}} \alpha = \lambda_2$ the second largest eigenvalue of $\bar{\bar{\Sigma}}$

The **second principal component** is the eigenvector associated with the second largest eigenvalue, λ_2

The total **variance is invariant** to a change of basis! $\sum_{i=1}^{d} \sigma_i^2 = \sum_{i=1}^{d} \lambda_i$

PCA algorithm

1. Normalize to zero-mean, unit-variance
2. Calculate the covariance matrix $\overline{\overline{\Sigma}}$ of the normalized data set
3. Find all eigenvalues of $\overline{\overline{\Sigma}}$ and arrange them in descending order

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d \geq 0$$

4. Choose *r* for the top *r* eigenvalues (and corresponding eigenvectors/principal components)

$\overline{\overline{U_r}} = \left[\overline{u_1}\,\overline{u_2} \ldots \overline{u_r}\right]$ the reduced basis (of principal components)

$\overline{a_i} = \overline{\overline{U_r}}^T \overline{x_i}, i = \overline{1,n}$ the reduced dimensionality data set

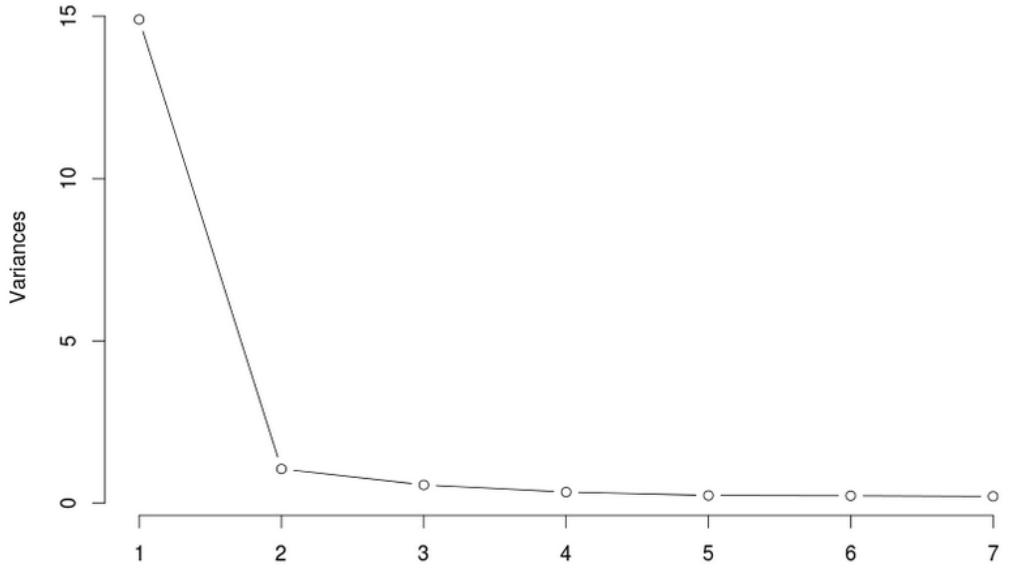
How to choose r?

$$r = \min\left\{ i \mid \frac{\lambda_1 + \ldots + \lambda_i}{\lambda_1 + \ldots + \lambda_i + \ldots \lambda_d} \geq \varepsilon, \varepsilon \in [0.7, 0.9] \right\}$$

Kaiser criterion (Henry Felix Kaiser, 1960) $r = \min\{ i \mid \lambda_i \geq 1 \}$

$$r = \min\{ i \mid \lambda_i \geq \frac{\lambda_1 + \ldots + \lambda_d}{d} \}$$

`screeplot`(*modelname*)



PCA Pros & Cons

PROS

- dimensionality reduction can help learning
- removes noise
- can deal with large data sets

CONS

- hard to interpret
- sample dependent
- linear (

Bibliography

- **A.A. Farag, S. Elhabian**, "*A Tutorial on Principal Component Analysis*" (2009) - http://dai.fmph.uniba.sk/courses/ml/sl/PCA.pdf
- **N. de Freitas**, "CPSC 340 - Machine Learning and Data Mining Lectures", University of British Columbia (2012) - http://www.cs.ubc.ca/~nando/340-2012/lectures.php
- **I. Georgescu**, *"Inteligență computațională"*, Editura ASE (2015)
- **I.T. Jolliffe**, "*Principal Component Analysis*", 2nd ed., Springer (2002)
- **J.N. Kutz**, "*Data-Driven Modeling & Scientific Computation. Methods for Complex Systems & Big Data*", Oxford University Press (2013)
- **J. N. Kutz**, "*Singular Value Decomposition*" - http://courses.washington.edu/am301/page1/page15/video.html
- **K. Pearson**, "*On lines and planes of closest to systems of points in space*", Philos. Mag, Sixth Series, 2, pp. 559-572 (1901) - http://stat.smmu.edu.cn/history/pearson1901.pdf
- **M.J. Zaki, W. Meira Jr.**, *"Data Mining and Analysis. Fundamental Concepts and Algorithms"*, Cambridge University Press (2014)