Predictive Analytics (ISE529)

Dimension Reduction (I)

Dr. Tao Ma
ma.tao@usc.edu

Tue/Thu, May 22 - July 1, 2025, Summer

USC Viterbi
School of Engineering
Daniel J. Epstein Department of Industrial and Systems Engineering

USC University of Southern California

PRINCIPAL COMPONENTS ANALYSIS

Dimension Reduction

- Assume we have a dataset represented in an $n \times D$ matrix X consisting of n data vectors x_i with dimensionality D. x_i is the ith row of X.
- Assume further that this dataset has intrinsic (inherent) dimensionality d (often $d \ll D$).
- Dimensionality reduction techniques transform dataset X with dimensionality D into a new dataset Y with dimensionality d, while retaining the geometry of the data as much as possible.
- Principal component analysis (PCA) constructs a low-dimensional representation of the data that describes as much of the variance in the data as possible.

Principal Components Analysis

The important properties of the new data set acquired by PCA:

- Principal components analysis transforms a set of observed variables into a new set of variables that are **uncorrelated** with one another.
- The transformed variables account for the variance of original variables in **sequentially decreasing proportions**.
- Each transformed variable in the new data set is called **PC score**. The first column PC score contains the greatest proportion of the total sample variance of the original data. The second column PC score accounts for a maximal proportion of the **remaining variance** subject to being uncorrelated with the first PC score. Subsequent components are defined similarly.
- The transformed variables (PC scores) are a linear combination of the original variables.

Spectral Decomposition

A. L. Cauchy established the Spectral Decomposition in 1829.

CAUCHY, A.L.(1789-1857)

Let A be a $m \times m$ real symmetric matrix. Then there exists an orthogonal matrix P such that $A = PDP^T$, where D is a diagonal matrix.

Spectral Decomposition

In multivariate analysis our data is a matrix. Suppose our data is matrix X. Suppose $X_{m\times n}$ is mean centered, i.e., $X \rightarrow (X-\mu)$

and the variance-covariance matrix is $\sum = (X\text{-}\mu)^T (X\text{-}\mu)$. The variance-covariance matrix $\sum$ is real and symmetric.

Using spectral decomposition, we can write $\sum = PDP^T$, where D is a diagonal matrix, i.e., $D = diag(\lambda_1, \lambda_2, \cdots, \lambda_n)$ and $\lambda_1, \ldots, \lambda_n$ are eigenvalues of $\sum$ *and* $\lambda_1 \geq \lambda_2 \geq, \cdots, \lambda_n$, P is a matrix where each column is the corresponding orthonormal eigenvector $u_1, \ldots, u_n$ *of* $\sum$. The first eigenvector column is the **first principal component**, the second eigenvector column is the **second principal component**, and so on.

$$\Sigma = \underbrace{\begin{pmatrix} \uparrow & \uparrow & & \uparrow \\ v_1 & v_2 & \cdots & v_n \\ \downarrow & \downarrow & & \downarrow \end{pmatrix}}_{P} \underbrace{\begin{pmatrix} \lambda_1 & & 0 \\ & \lambda_2 & \\ & & \ddots \\ 0 & & \lambda_n \end{pmatrix}}_{\Lambda} \underbrace{\begin{pmatrix} \leftarrow & v_1 & \rightarrow \\ \leftarrow & v_2 & \rightarrow \\ & \vdots & \\ \leftarrow & v_n & \rightarrow \end{pmatrix}}_{P^T}$$

Spectral Decomposition

The new data (PC scores) via PC transformation is:

$$\boldsymbol{Y = (X-\mu)\, P}$$

where,

$E(Y_i) = 0$

$Var(Y_i) = \lambda_i$

$Cov(Y_i, Y_j) = 0$ if $i \neq j$

$Var(Y_1) \geq Var(Y_2) \geq \ldots \geq Var(Y_n)$

$$\sum_{i=1}^{n} Var(Y_i) = tr(\Sigma) = \text{Total variation of Data} = tr(D)$$

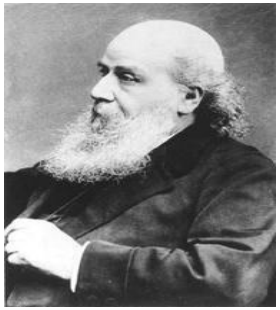
$$\prod_{i=1}^{n} Var(Y_i) = |\Sigma|$$

Singular Value Decomposition (SVD)

There are five mathematicians who made great contributions to establishment of the singular value decomposition and developing its theory, among others.

Eugenio Beltrami (1835-1899)

Camille Jordan (1838-1921)



James Joseph Sylvester (1814-1897)

Erhard Schmidt (1876-1959)

Hermann Weyl (1885-1955)

The Singular Value Decomposition was originally developed by Eugenio Beltrami and Camille Jordan two mathematician in the mid to late 1800's.

Several other mathematicians took part in the final developments of the SVD including James Joseph Sylvester, Erhard Schmidt and Hermann Weyl who studied the SVD until the mid-1900's.

C.Eckart and G. Young prove low rank approximation of SVD (1936).

What is SVD?

Any real ($m \times n$) matrix X, where ($n \leq m$), can be decomposed,

$$X = U \Lambda V^T$$

- U is a ($m \times n$) column orthonormal matrix ($U^T U = I$), containing the eigenvectors of the symmetric matrix XX^T.
- Λ is a ($n \times n$) diagonal matrix, containing the **singular values** of matrix X. The number of non-zero diagonal elements of Λ corresponds to the rank of X.
- V^T is a ($n \times n$) row orthonormal matrix ($V^T V = I$), containing the eigenvectors of the symmetric matrix $X^T X$.

What is SVD?

Any real ($m \times n$) rectangular matrix X, where ($n \leq m$), can be decomposed,

$$X = U \Lambda V^T$$

$$X_{m \times n} = \underbrace{\begin{pmatrix} \uparrow & \uparrow & & \uparrow \\ u_1 & u_2 & \cdots & u_n \\ \downarrow & \downarrow & & \downarrow \end{pmatrix}}_{U_{m \times n}} \underbrace{\begin{pmatrix} \sqrt{\lambda_1} & & & 0 \\ & \sqrt{\lambda_2} & & \\ & & \ddots & \\ 0 & & & \sqrt{\lambda_n} \end{pmatrix}}_{\Lambda_{n \times n}} \underbrace{\begin{pmatrix} \leftarrow & v_1 & \rightarrow \\ \leftarrow & v_2 & \rightarrow \\ & \vdots & \\ \leftarrow & v_n & \rightarrow \end{pmatrix}}_{V^T_{n \times n}}$$

Note that "singular values" of a matrix X are the positive **square roots of the eigenvalues** of the matrix product XX^T or X^TX, while "eigenvalues" of a matrix are simply the characteristic roots of the matrix itself,

SVD Based PCA

Suppose X is a mean centered data matrix, Then X using SVD,

$$X = U\Lambda V^{\mathrm{T}}$$

The new data (PC scores) can be calculated with

$$\boldsymbol{Y = XV = U\Lambda}$$

Then the first columns of Y represents the first principal component score and so on.

- SVD Based PC is more Numerically Stable.
- If no. of variables is greater than no. of observations (m<<n), then SVD based PCA will give efficient result

Spectral Decomposition vs. SVD

Suppose X is a mean centered data matrix, Then X using SVD,

$$X = U\Lambda V^{\mathrm{T}}$$

The variance-covariance matrix of X can be written as

$$\begin{aligned}\sum = X^T X &= (U\Lambda V^T)^T\, U\Lambda V^T \\ &= V\Lambda U^T U\Lambda V^T\end{aligned}$$

Because U is orthornormal eigenvector matrix, then $U^T U = I$

Hence,

$$\begin{aligned}\sum = X^T X &= (U\Lambda V^T)^T\, U\Lambda V^T \\ &= V\Lambda U^T U\Lambda V^T \\ &= V\Lambda\Lambda V^T \\ &= VDV^T = PDP^{\mathrm{T}}\end{aligned}$$

where $D = \Lambda^2$, $V = P$

Variance of New Data (PC Scores)

SVD based Principal Component Analysis (PCA) constructs a low dimensional representation of the data that describes as much of the variance in the data as possible.

The new data set (PC scores) can be written as

$$Y = XV = V^T X = UA$$

where ***X* is mean centered (mean is removed)**. The variance-covariance of PC score variables Y is

$$\begin{aligned} \text{Var}(Y) &= \text{E}\,(XV)^T(XV) \\ &= V^T X^T XV \\ &= V^T \sum V \\ &= V^T PDP^T V \\ &= D = diag(\lambda_1, \lambda_2, \cdots, \lambda_n) \end{aligned}$$

where $V = P$, *and* $V^T P = I$.

Principal Components Analysis

Why do we need principal component technique?

1. Dimension reduction

The general hope of principal components analysis is that the first few components will account for a substantial proportion of the variation in the original variables, $x_1, \ldots, x_q$, and can, consequently, be used to provide a convenient **lower-dimensional summary of these variables** that might prove useful for a variety of reasons.

2. As uncorrelated input to some other analysis, such as regression analysis

- There are **too many** explanatory variables relative to the number of observations.
- The explanatory variables are highly **correlated**.

Eigenvalues and Eigenvectors

How are the principal components found?

The first PC is the **eigenvector of the sample covariance matrix**, $\sum$, corresponding to this matrix's largest eigenvalue. The second PC is the eigenvector of the sample covariance matrix, $\sum$, corresponding to this matrix's second largest eigenvalue, and so on.

Recall the characteristic equation

$$\Sigma\gamma = \lambda\gamma$$

λ are eigenvalues, γ eigenvectors, and $\sum$ is a ***square*** matrix. The eigenvalues can be found by solving

$$\det(\sum - \lambda I) = 0$$

where I is the identity matrix. Then the eigenvectors can be found by solving

$$(\sum - \lambda I)\gamma = 0$$

Orthogonal and Orthonormal vectors

Definition. We say that 2 vectors are orthogonal if they are perpendicular to each other. i.e., the dot product of the two vectors is zero.

Example: the set of vectors is mutually orthogonal. $\begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 \\ \sqrt{2} \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ -\sqrt{2} \\ 1 \end{pmatrix}$

$$\begin{aligned} (1,0,-1).(1,\sqrt{2},1) &= 0 \\ (1,0,-1).(1,-\sqrt{2},1) &= 0 \\ (1,\sqrt{2},1).(1,-\sqrt{2},1) &= 0 \end{aligned}$$

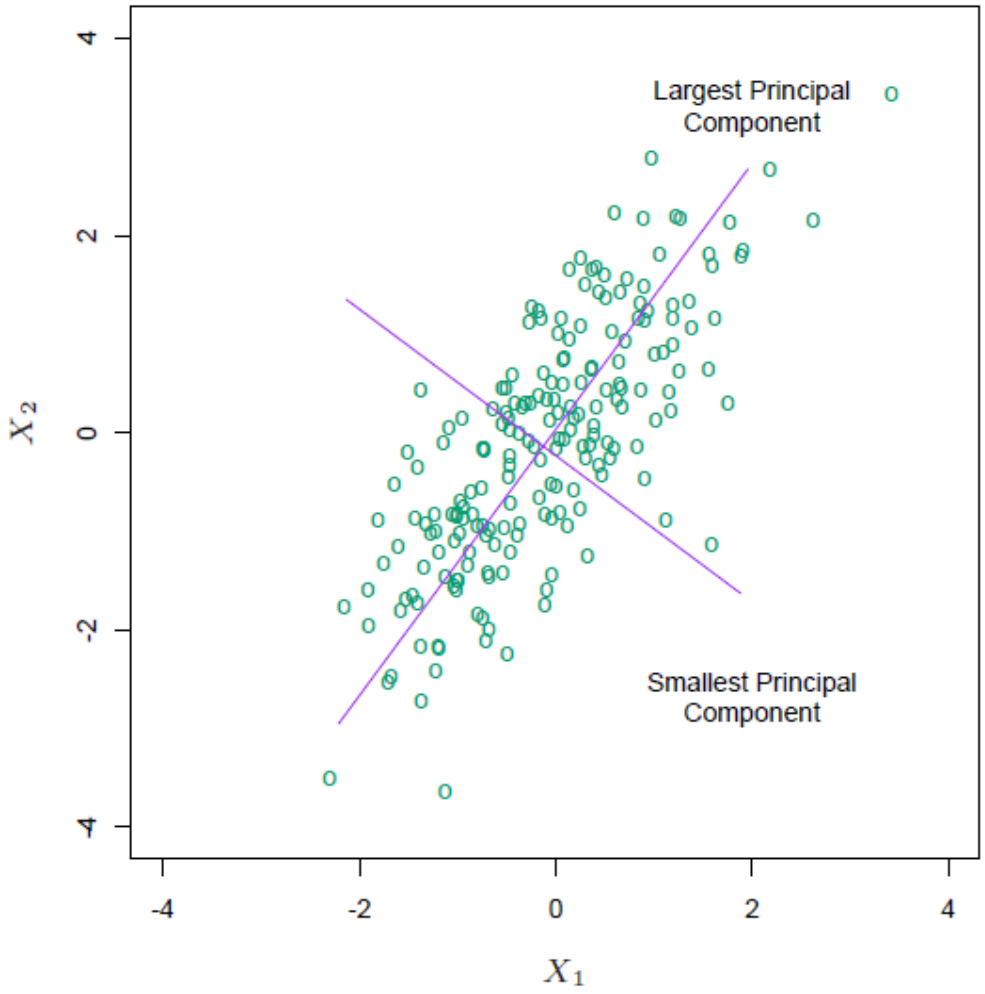
Definition. A set of vectors *S* is orthonormal if every vector in *S* has norm 1 (unit vector) and the set of vectors are mutually orthogonal.

Example: let

$$\begin{aligned} \vec{u}_1 &= \frac{\vec{v}_1}{|\vec{v}_1|} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} = \begin{pmatrix} 1/\sqrt{2} \\ 0 \\ -1/\sqrt{2} \end{pmatrix} \\ \vec{u}_2 &= \frac{\vec{v}_2}{|\vec{v}_2|} = \frac{1}{2} \begin{pmatrix} 1 \\ \sqrt{2} \\ 1 \end{pmatrix} = \begin{pmatrix} 1/2 \\ \sqrt{2}/2 \\ 1/2 \end{pmatrix} \\ \vec{u}_3 &= \frac{\vec{v}_3}{|\vec{v}_3|} = \frac{1}{2} \begin{pmatrix} 1 \\ -\sqrt{2} \\ 1 \end{pmatrix} = \begin{pmatrix} 1/2 \\ -\sqrt{2}/2 \\ 1/2 \end{pmatrix} \end{aligned}$$

The set of vectors $\{\vec{u}_1, \vec{u}_2, \vec{u}_3\}$ is orthonormal.

PC1 vs. PC2



The figure shows that PC1 vs. PC2 principal components of some data points in two dimensions.

Total Variance

The total variance of the ***q*** principal components will equal the total variance of the original variables so that

$$\sum_{i=1}^{q} \lambda_i = s_1^2 + s_2^2 + \cdots + s_q^2,$$

where S_i^2 is the sample variance of x_i, the diagonal element of the sample covariance matrix $\sum$. We can write this more concisely as

$$\sum_{i=1}^{q} Var(Y_i) = \sum_{i=1}^{q} \lambda_i = tr(\Sigma)$$

$$= \text{Total variation of Data} = tr(D)$$

Account for Total Variance

Consequently, the *j*th principal component accounts for a proportion P_j of the total variation of the original data, where

$$P_j = \frac{\lambda_j}{trace(\Sigma)}$$

The first m principal components, where $m < q$ account for a proportion $P^{(m)}$ of the total variation in the original data, where

$$P^{(m)} = \frac{\sum_{j=1}^{m} \lambda_j}{trace(\Sigma)}$$

Determine the Number of PCs

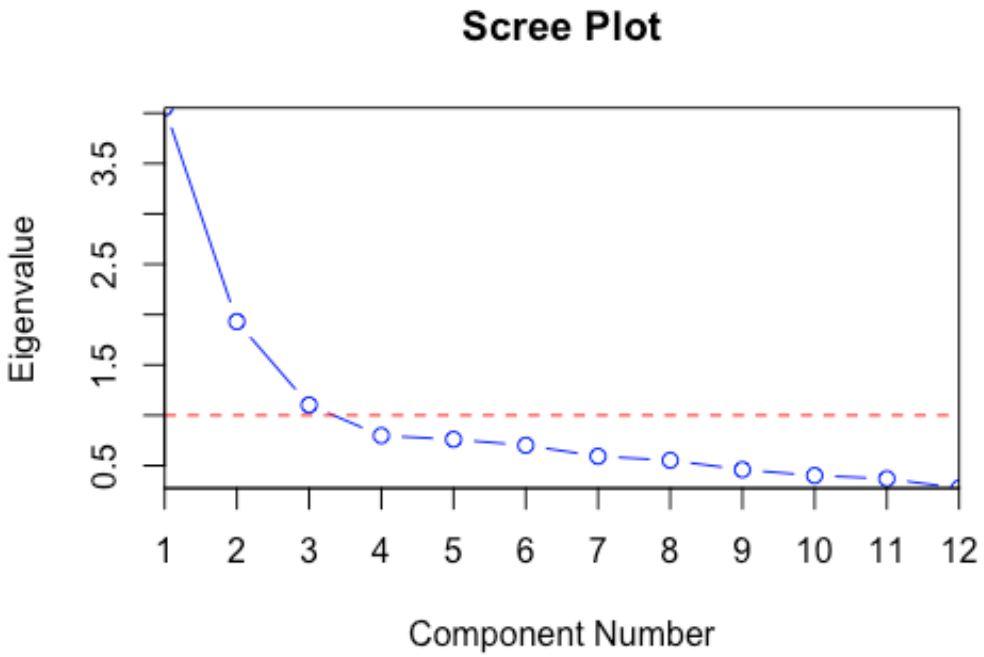
How many components are needed to provide an adequate summary of a given data set? The most common of the relatively ad hoc procedures that have been suggested.

- Retain just enough components to explain some specified large percentage of the total variation of the original variables. Values between 70% and 90% are usually suggested, although smaller values might be appropriate as ***q*** or ***n***, the sample size, increases.

- Exclude those principal components whose eigenvalues are less than the average, $\sum_{i=1}^{q} \lambda_i / q$. Since $\sum_{i=1}^{q} \lambda_i = trace(\Sigma)$, the average eigenvalue is also the average variance of the original variables. This method then retains those components that account for more variance than the average for the observed variables

Determine the Number of PCs

- Examination of the plot of the λ_i against *i*, the so–called scree diagram. The number of components selected is the value of *i* corresponding to an "*elbow*" in the curve, i.e., a change of slope from "steep" to "shallow".



- A modification of the scree diagram is the log–eigenvalue diagram consisting of a plot of log(λ_i) against *i*.

Covariance vs. Correlation Matrix

Should principal components be extracted from the covariance or the correlation matrix?

One problem with principal components analysis is that it is not scale-invariant. Suppose the three variables in a multivariate data set are weight in **pounds**, height in **feet**, and age in **years**, but for some reason we would like our principal components expressed in **ounces**, **inches**, and **decades**. Intuitively two approaches seem feasible;

1. Multiply the variables by 16, 12, and 1/10, respectively and then carry out a principal components analysis on the covariance matrix of the three variables.

2. Carry out a principal components analysis on the covariance matrix of the original variables and then multiply the elements of the relevant component by 16, 12, and 1/10.

Unfortunately, these two procedures do not generally lead to the same result.

Covariance vs. Correlation Matrix

Principal components should only be extracted from the sample **covariance matrix, S,** when all the original variables have **roughly the same scale**. But this is rare in practice and consequently, in practice, principal components are extracted from the **correlation matrix** of the variables, **R**.

Extracting the components as the eigenvectors of **R** is equivalent to calculating the principal components from the original variables after each has been standardized to have **unit variance**. It should be noted, however, that there is rarely any simple correspondence between the components derived from **S** and those derived from **R**. And choosing to work with **R** rather than with **S** involves a definite but possibly arbitrary decision to make variables "**equally important**".

PRINCIPAL COMPONENTS REGRESSION

Principal Components Regression

- PCR dimension reduction methods work in two steps. First, the transformed predictors $Z_1, Z_2, \ldots, Z_M$ are obtained. Second, the model is fit using these M predictors with least squares.

- When performing PCR, we generally recommend standardizing each predictor prior to generating the principal components. This standardization ensures that all variables are on the same scale.

- In the absence of standardization, the high-variance variables will tend to play a larger role in the principal components obtained, and the scale on which the variables are measured will ultimately influence the final PCR model.

Principal Components Regression

In many situations we have a large number of inputs, often very correlated. The PCR method produces a small number of linear combinations $\boldsymbol{Z}_m$, $m = 1, \ldots, M$ of the original inputs $\boldsymbol{X}_j$, and the $\boldsymbol{Z}_m$ are then used in place of the $\boldsymbol{X}_j$ as inputs in the regression. In this approach the linear combinations $\boldsymbol{Z}_m$ used are the principal component **scores** defined as

$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j$$

where $\phi_m = [\phi_{1m}, \phi_{2m} \ldots, \phi_{pm}]^T$ is the mth eigenvector, $m = 1, \ldots, M$.

Principal Components Regression

and then regresses $\boldsymbol{y}$ on $z_1, z_2, \ldots, z_M$ using least squares for some $M \leq p$. Since the z_m are orthogonal, this regression is just a sum of univariate regressions:

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \epsilon_i, \quad i = 1, \ldots, n,$$

where $\theta_0, \theta_1, \ldots, \theta_M$ are the regression coefficients, which can be estimated by

$$\hat{\theta}_m = \langle \mathbf{z}_m, \mathbf{y} \rangle / \langle \mathbf{z}_m, \mathbf{z}_m \rangle$$

Principal Components Regression

Since the z_m are each linear combinations of the original x_j, we can express the solution in terms of coefficients of the x_j

$$\sum_{m=1}^{M} \theta_m z_{im} = \sum_{m=1}^{M} \theta_m \sum_{j=1}^{p} \phi_{jm} x_{ij} = \sum_{j=1}^{p} \sum_{m=1}^{M} \theta_m \phi_{jm} x_{ij} = \sum_{j=1}^{p} \beta_j x_{ij},$$

where

$$\beta_j = \sum_{m=1}^{M} \theta_m \phi_{jm}.$$

The term dimension reduction comes from the fact that this approach reduces the problem of estimating the (p+1) coefficients $\beta_0, \beta_1, \ldots, \beta_p$ to the simpler problem of estimating the (M+1) coefficients $\theta_0, \theta_1, \ldots, \theta_M$, where M < p.

PC Line vs. Least Squared Line

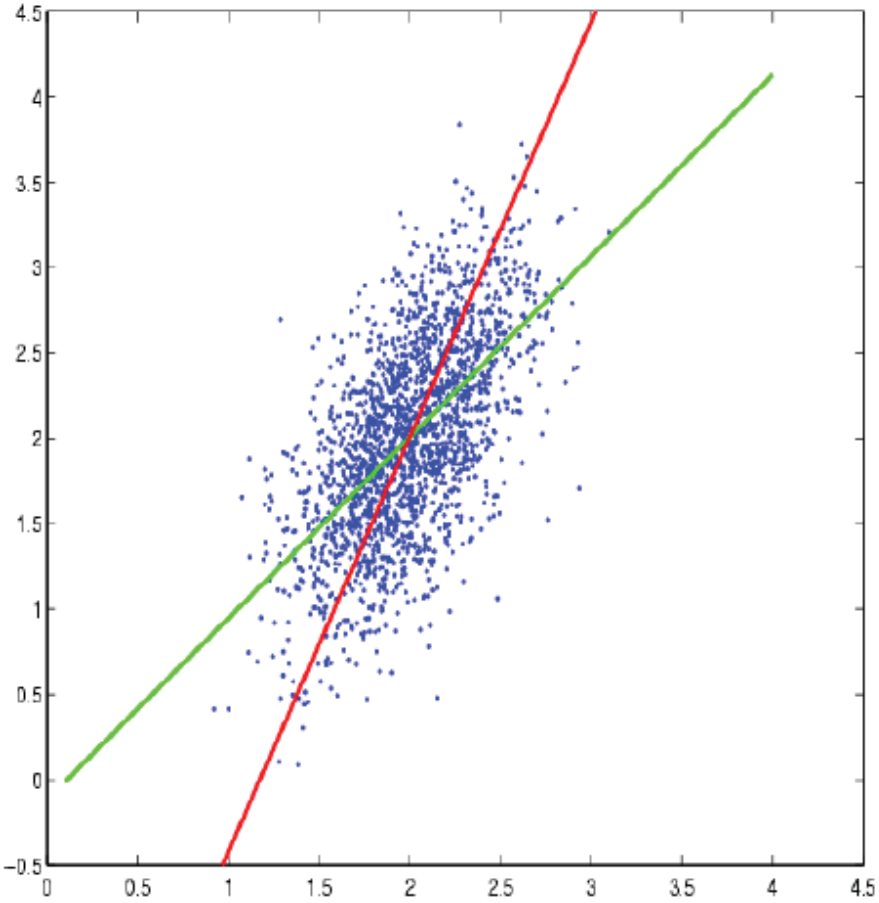


Figure: The least squares line minimizes the vertical squared distance, but the 1st PC line minimized the perpendicular squared distance.