Principal Components Analysis

Suppose I observe a dataset consisting of n observations of p features:

$$\{(x_{i1}, x_{i2}, \ldots, x_{ip})\}_{i=1}^{n}.$$

We can arrange these in an $n \times p$ design matrix:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} - & x_1 & - \\ - & x_2 & - \\ & \vdots & \\ - & x_n & - \end{bmatrix}.$$

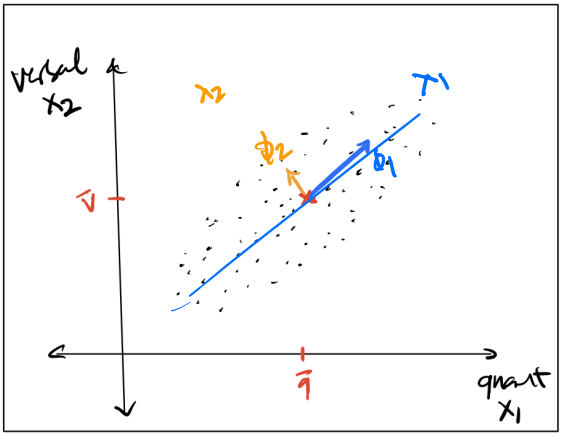
Let's call the i^{th} row $x_i \in \mathbb{R}^p$, & let $\bar{x}_j \in \mathbb{R}$ denote the mean of the j^{th} column:

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}.$$

Then let $\bar{x} \in \mathbb{R}^p$ be the column vector w/ elements $\bar{x} = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p)^T$. The main idea of principal components analysis (PCA) is that sometimes the coordinate system we get from our features, i.e. the original $(x_1, x_2, \ldots, x_p)$-axes, is not the coordinate system which agrees or aligns most with the natural variation in our data.

Here's a simple example in $\mathbb{R}^2$:

Ex: Suppose n children took a standardized exam which had a quantitative subscore x_1, & a verbal subscore x_2. The data $\{(x_{i1}, x_{i2})\}_{i=1}^{n}$ are distributed roughly like this:



Despite the data coming to us in the form of quant & verbal scores, the directions in which the data vary seem to indicate other unobserved factors or "components." In this example, we might call them something like:

1) general aptitude as measured by the test &
2) the difference between quantitative & verbal aptitude.

Steps for Conducting PCA

Principal components analysis attempts to find these directions/new coordinates through the following steps:

1. Center each column vector to create the centered matrix

$$X^c = \{x_{ij}^c\}_{i=1,j=1}^{n,p}$$

where $x_{ij}^c = x_{ij} - \bar{x}_j$

Visually, this corresponds to moving the origin to the mean point of the data cloud. (The red X.)

2. Calculate the sample covariance matrix

$$S = \frac{1}{n-1} X^{cT} X^c. \quad = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T$$

Notice this is a symmetric, positive semidefinite matrix, so we can...

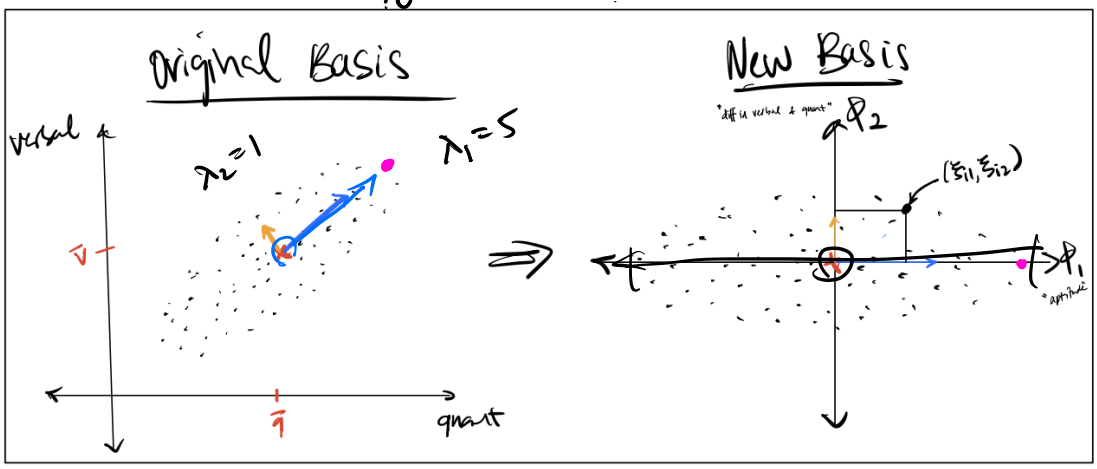
3. Calculate the eigendecomposition:

$$S = \Phi \Lambda \Phi^T \quad \text{(AKA spectral decomposition).}$$

The cols of $\Phi = [\phi_1 \ \phi_2 \ \dots \ \phi_p]$ give us the unit vectors in the directions of our new coordinate system.

Here Λ is a diagonal matrix of the eigenvalues, that is,

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{bmatrix}.$$



$x_i = \bar{x} + 5\phi_1 + 0\phi_2$

$\xi_{i1} = +5$

$\xi_{i2} = 0$

$\langle x_i - \bar{x}, \phi_1 \rangle = \xi_{i1} \rightarrow \quad \langle x_i - \bar{x}, \phi_2 \rangle = \xi_{i2}$

4. Once the new directions/unit vectors $\phi_1, \ldots, \phi_p$ have been chosen, how can we compute the coordinates of a datapoint x_i with respect to our new basis? In other words, how can we find coefficients $\xi_{i1}, \ldots, \xi_{ip}$ such that

$$x_i \;=\; \bar{x} + \xi_{i1}\phi_1 + \ldots + \xi_{ip}\phi_p?$$

We can compute these coefficients using *Fourier's trick*. Centering and taking the dot product of both sides of the equation,

$$x_i - \bar{x} \;=\; \xi_{i1}\,\phi_1 \;+\; \cdots \;+\; \xi_{ip}\,\phi_p \tag{1}$$

with the vector ϕ_1, we obtain the coordinate in the first direction,

$$\langle x_i - \bar{x}, \phi_1 \rangle \;=\; \xi_{i1}\underbrace{\langle \phi_1, \phi_1 \rangle}_{=1} + \xi_{i2}\underbrace{\langle \phi_2, \phi_1 \rangle}_{=0} + \cdots + \xi_{ip}\underbrace{\langle \phi_p, \phi_1 \rangle}_{=0} \implies \xi_{i1} = \langle x_i - \bar{x}, \phi_1 \rangle,$$

and so on for $\xi_{i2} = \langle x_i - \bar{x}, \phi_2 \rangle$, ... , $\xi_{ip} = \langle x_i - \bar{x}, \phi_p \rangle$, $\quad i = 1, ..., n.$

Facts about PCA:

1. The coordinates of the i^{th} data point in the new basis are given by the inner product of the i^{th} row of X^c (i.e. $x_i^C \in \mathbb{R}^p$)) & the directions $\phi_1, \phi_2, \ldots, \phi_p$:

$$\xi_{i1} = \langle x_i - \bar{x}, \phi_1 \rangle$$

$$\vdots$$

$$\xi_{ip} = \langle x_i - \bar{x}, \phi_p \rangle$$

$\text{proj}_v(u) = \frac{\langle u, v \rangle}{\|v\|^2} = \langle u, v \rangle$, if $\|v\|^2 = 1$

These are called the *principal component scores.*
They express the original data in terms of the new basis:

$$x_i = \bar{x} + \xi_{i1}\phi_1 + \xi_{i2}\phi_2 + \cdots + \xi_{ip}\phi_p$$

2. The sample variance of the j^{th} PC score is equal to the j^{th} eigenvalue of S:

$$\frac{1}{n-1}\sum_{i=1}^{n}(\xi_{ij} - \bar{\xi}_j)^2 \;=\; \frac{1}{n-1}\sum_{i=1}^{n}\xi_{ij}^2 = \lambda_j, \quad \forall j = 1, \ldots, p.$$

3. Because the spectral decomposition arranges the eigenvalues along the diagonal in decreasing order, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$, the basis vector ϕ_1 corresponds to the direction in which the data varies the most, followed by ϕ_2, then ϕ_3, & so on.

4. If we perform SVD on $X^c = UDV^T$, then the columns of V are the same as the eigenvectors of S. i.e. $\phi_j = v_j$ & $j = 1, \ldots, p$, and the rows of UD correspond to the coordinates of the original data points projected into principal component space.

$$UD = \begin{bmatrix} \xi_{11} & \xi_{12} & \cdots & \xi_{1p} \\ \xi_{21} & \xi_{22} & & \xi_{2p} \\ \vdots & & & \end{bmatrix}$$

5. The "total variation" of X (the summed variance of each of the p features) is equal to the trace of S:

$$\sum_{j=1}^{p}\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\bar{x}_{\cdot j})^2\right] = \operatorname{tr}(S) = \lambda_1+\lambda_2+\cdots+\lambda_p,$$

& the ratio

$$\frac{\lambda_k}{\lambda_1+\lambda_2+\cdots+\lambda_p}$$

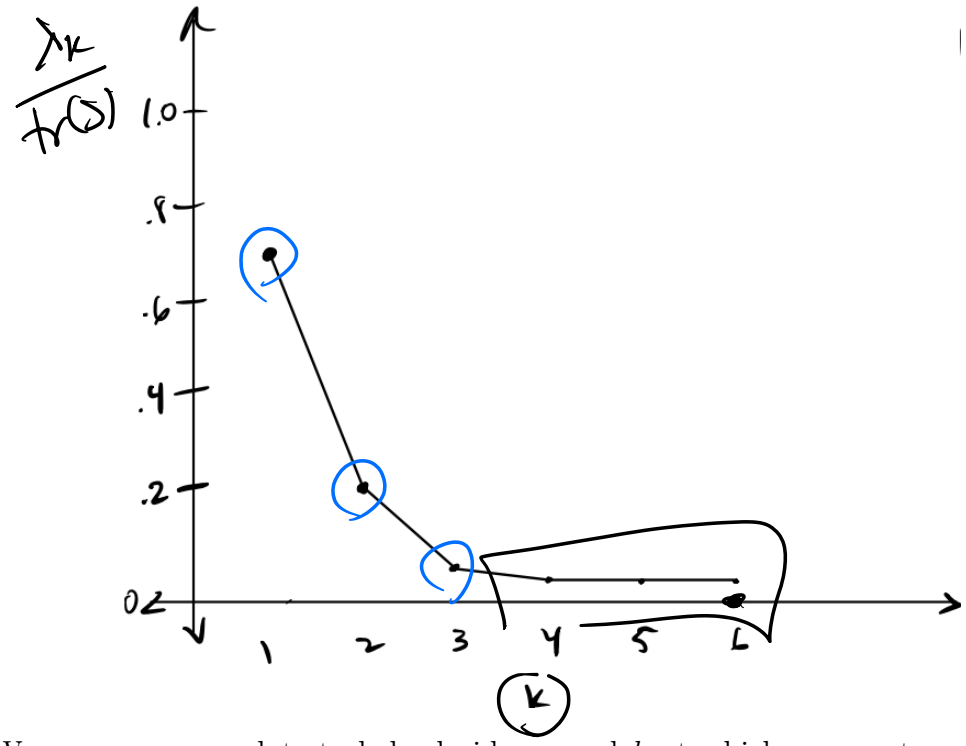
estimates the total variation in X that is explained by the k^{th} principal component.

6. Often we plot

$$\frac{\lambda_k}{\operatorname{tr}(S)}$$

against k which shows how the eigenvalues decay. This is called a *scree plot.*

Ex: if $x_i \in \mathbb{R}^6$ we may have a scree plot like:



7. We can use scree plots to help decide a good k at which we can truncate the basis expansion of x_i without losing much variation/information:

$$x_i \approx \bar{x} + \sum_{j=1}^{k}\xi_{ij}\phi_j \quad \text{(rank } k \text{ approximation)}$$

(2) Why?

$$\frac{1}{n-1}\sum_{i=1}^{n} \xi_{ij}^2 = \frac{1}{n-1}\sum_{i=1}^{n} \langle x_i - \bar{x}, \phi_j \rangle^2$$

$$= \frac{1}{n-1}\sum_{i=1}^{n} \left[(x_i-\bar{x})^T \phi_j\right]^2$$

$$= \frac{1}{n-1}\sum_{i=1}^{n} \left[(\phi_j^T (x_i-\bar{x})\right]\left[(x_i-\bar{x})^T \phi_j\right]$$

$$= \frac{1}{n-1}\,\phi_j^T \left[\sum_{i=1}^{n} (x_i-\bar{x})(x_i-\bar{x})^T\right]\phi_j$$

$$= \phi_j^T S \phi_j = \phi_j^T (\lambda_j \phi_j)$$

$$= \lambda_j \phi_j^T \phi_j = \lambda_j \|\phi_j\|^2$$

$$= \lambda_j \cdot 1 = \lambda_j$$

4. $X^c = UDV^T$

$$S = \frac{1}{n-1} X^{cT} X^c$$

$$= \frac{1}{n-1} (UDV^T)^T (UDV^T)$$

$$= \frac{1}{n-1} (V^T D^T U^T U D V^T) = \frac{1}{n-1} (V D \cdot I \cdot D V^T$$

$$= \frac{1}{n-1} (V D^2 V^T)$$

$$= V \left(\frac{D^2}{n-1}\right) V^T \Rightarrow \Phi = V$$

or

$\phi_1 = v_1$

$\phi_2 = v_2 \ldots$

$\Phi \Lambda \Phi^T$

$$S = V \left(\frac{D^2}{n-1}\right) V^T$$

$$V^T S V = V^T V \left(\frac{D^2}{n-1}\right) V^T V$$

$$V^T \left(\lambda_1 v_1 \cdots \lambda_p v_p\right) = D^2/n-1$$

$$\begin{pmatrix} v_1^T \\ \vdots \\ v_p^T \end{pmatrix} \left(\lambda_1 v_1 \cdots \lambda_p v_p\right)$$

$$= \begin{pmatrix} \lambda_1 & & 0 \\ & \lambda_2 & \\ 0 & & \lambda_p \end{pmatrix} = D^2/n-1$$

(S) $$S = \begin{bmatrix} Var(X_1) & Cov(X_1,X_2) \cdots & Cov(X_1,X_p) \\ & Var(X_2) & \\ & & \ddots \\ & & Var(X_p) \end{bmatrix}$$

$$tr(S) = Var(X_1) + Var(X_2) \cdots + Var(X_p)$$

$$= \sum_{j=1}^{p} \left(\frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2\right)$$

$tr(ABC) = tr(BCA)$ ← "cyclic"

$$tr(S) = tr(\Phi \Lambda \Phi^T) = tr(\Lambda \Phi^T \Phi)$$

$$= tr(\Lambda I)$$

$$= tr(\Lambda) = \lambda_1 + \lambda_2 \cdots + \lambda_p$$

Why is UD the coordinates in the new basis?

$$X^C = UDV^T$$

$$X^C V = UDV^T V = UD$$

$$\begin{bmatrix} x_1 - \bar{x} \\ x_2 - \bar{x} \\ \vdots \\ x_p - \bar{x} \end{bmatrix} \begin{bmatrix} | & & | \\ v_1 & \cdots & v_p \\ | & & | \end{bmatrix} = UD$$

$$= \begin{bmatrix} \langle x_1 - \bar{x}, v_1 \rangle & \langle x_1 - \bar{x}, v_2 \rangle & \cdots & \langle x_1 - \bar{x}, v_p \rangle \\ & \vdots & & \\ \langle x_n - \bar{x}, v_1 \rangle & \cdots & & \langle x_n - \bar{x}, v_p \rangle \end{bmatrix} = UD$$

(v_1, v_2, v_p are the principal components ϕ_1, ϕ_2, ϕ_p)

$$= \begin{bmatrix} \xi_{11} & \xi_{12} & \cdots & \xi_{1p} \\ \vdots & & & \\ \xi_{n1} & \xi_{n2} & \cdots & \xi_{np} \end{bmatrix} = UD$$

$$x_i = \bar{x} + \xi_{i1}\phi_1 + \xi_{i2}\phi_2 + \cdots + \xi_{ip}\phi_p$$

$$\approx \bar{x} + \xi_{i1}\phi + \xi_{i2}\phi_2 + \cdots + \xi_{ik}\phi_{ik} + 0 + \cdots + 0$$

if $\lambda_{k+1} \approx 0$

$\lambda_{k+2} \approx 0$

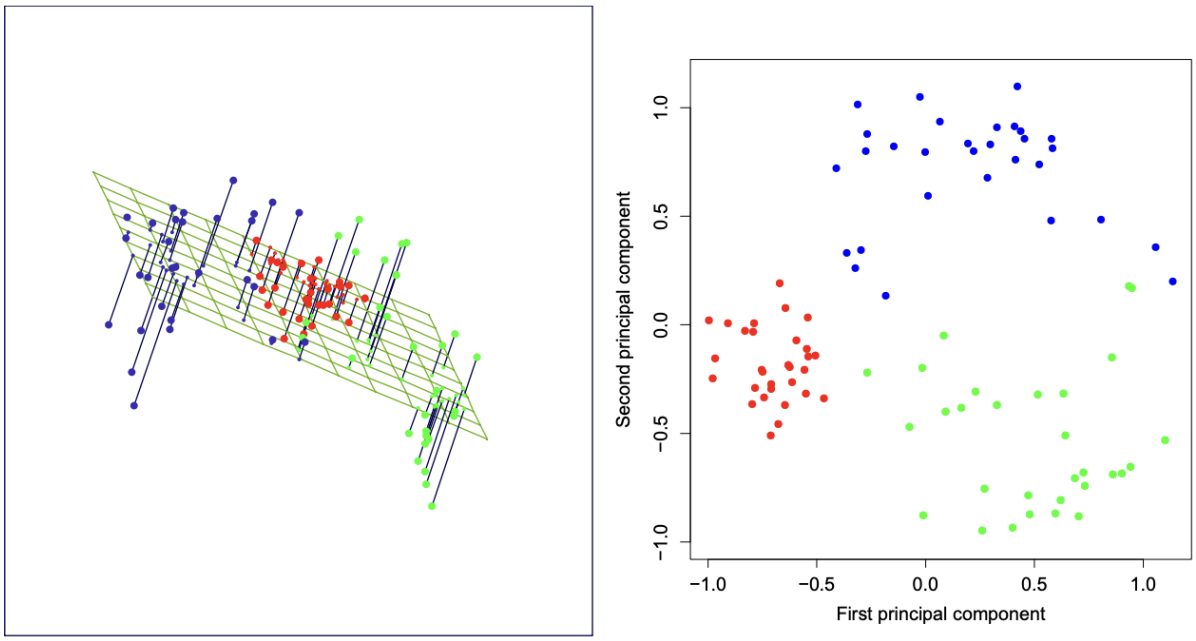
$\vdots$

$\lambda_p \approx 0$

↑ rank k approximation of the i^{th} data pt.

Another example for data in $\mathbb{R}^3$:

Consider a dataset consisting of 3-dimensional vectors whose coordinates are noise-contaminated position measurements on a sphere (left). Principal components analysis gives us the optimal low dimensional approximation (here, 2-dimensional) to the sphere-generated data under linear projection. The right panel shows the projection of the data onto the first two principal components, where the coordinates in PC space are given by the rows of UD in the SVD of the centered data matrix.



Recreated from *Elements of Statistical Learning*.

Computation Ex:

Given a design/data matrix

$$X = \begin{bmatrix} 6 & -4 \\ -3 & 5 \\ -2 & 6 \\ 7 & -3 \end{bmatrix}$$

Let's use PCA to approximate the original 4 datapoints by their rank-1 approximations.

Steps:

1. Center the data.

$$X^c = \begin{bmatrix} 4 & -5 \\ -5 & 4 \\ -4 & 5 \\ 5 & -4 \end{bmatrix}$$

2. Calculate

$$S = \frac{1}{3} X^{cT} X^c = \tfrac{1}{3} \begin{bmatrix} 82 & -80 \\ -80 & 82 \end{bmatrix}$$

$$\Rightarrow \text{eigenvectors} : \phi_1 = \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix}, \quad \phi_2 = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$$

$$\lambda_1 = 54 \quad \lambda_2 = 7/3$$

3. Project onto ϕ_1 for the scores:

$$\xi_{11} = [4 \ -5] \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix} = 9/\sqrt{2}$$

$$\xi_{21} = [-5 \ 4] \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix} = -9/\sqrt{2}$$

$$\xi_{31} = [-4 \ 5] \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix} = 9/\sqrt{2}$$

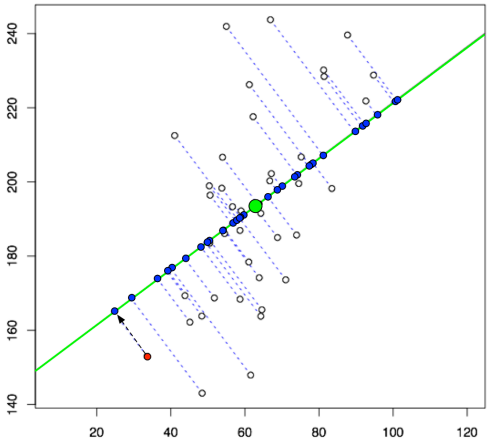
$$\xi_{41} = [5 \ -4] \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix} = 9/\sqrt{2}$$

4. Estimate x_1 by its rank 1 approximation.

$$\begin{bmatrix} 9 \\ 4 \end{bmatrix} = x_1 \approx \bar{X} + \frac{9}{\sqrt{2}} \phi_1 = \begin{bmatrix} 2 \\ 2 \end{bmatrix} + \frac{9}{\sqrt{2}} \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix} = \begin{bmatrix} 2 + 9/2 \\ 2 - 9/2 \end{bmatrix} = \begin{bmatrix} 6.5 \\ -2.5 \end{bmatrix}$$

Seeing the connection to SVD

We can think about PCA as searching for the optimal basis whose first directions align maximally with the variation in our data. We can explore this idea by noticing that if the first axis of our coordinate system is well-aligned with the data, then the PC scores $\xi_{11}, ..., \xi_{n1}$, will have a high sample variance.



In this example, the first axis of our coordinate system is as well-aligned with the data as possible. As a result, the points in blue (the first PC scores, i.e. projections onto the first direction) are very spread out. This means that the numbers $\xi_{11}, ..., \xi_{n1}$ have a maximally high sample variance.

Mathematically to find this first direction that aligns most with the data, we are essentially searching for the for the unit vector w that maximizes the sample variance of the first PC scores. In other words, we search for w with unit norm that maximizes

$$\begin{aligned}
\frac{1}{n-1}\sum_{i=1}^{n}\xi_{i1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\langle x_i - \bar{x},\, w\rangle^2 &= \frac{1}{n-1}\sum_{i=1}^{n} w^T(x_i-\bar{x})(x_i-\bar{x})^T w \\
&= \frac{1}{n-1} w^T\Big(\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})^T\Big) w \\
&= \frac{1}{n-1} w^T X^{cT} X^c\, w \\
&= \frac{1}{n-1}\|X^c\, w\|^2,
\end{aligned}$$

where X^C is the $n \times p$ matrix whose ith row is $(x_i - \bar{x})^T$. Equivalently, we choose ϕ_1 to be the unit vector w that maximizes $\|X^c\, w\|$. But this is precisely the optimization problem that is solved by the singular value decomposition. If

$$X^c = UDV^T$$

is a singular value decomposition of X^c, then we can take v_1 to be the first column of V.

Next to find the second direction, we choose ϕ_2 to be the unit vector which maximizes $\|X^c\, w\|$ subject to the constraint that w is orthogonal to ϕ_1. But again, this optimization problem is solved by the SVD. We can take ϕ_2 to be the second column of V.

The remaining vectors $\phi_3, \ldots, \phi_p$ are chosen in the same way: ϕ_j is chosen to be the unit vector w which maximizes $\|X^c w\|$ subject to the constraints that phi_j must be orthogonal to each of the vectors $\phi_1, \ldots, \phi_{j-1}$. Again, this is precisely the optimization problem solved by the SVD. We can take ϕ_j to be the jth column of V. So, by computing the SVD of the matrix X^c, we have found the principal component vectors $\phi_1, \ldots, \phi_p$.