Principal Component Analysis

PCA

Real world data and information therein may be:

- **Redundant**
  - One variables may carry the same information as the other variable
  - Information covered by a set of variable may overlap
- **Noisy**
  - Some dimensions may not carry any useful information and the variation in that dimension is purely due to noise in the observations

Important questions:

- how to reduce the dimensionality of the data
- what is the intrinsic dimensionality of the data?

PCA

- A principle component analysis is concerned with explaining the variance-covariance structure of a set of variables through a few linear combinations of these variables.
- Although p components are required to reproduce the total system variability, often much of this variability can be accounted for by a small number k of the principle components.

let the random vector $\boldsymbol{X}' = [X_1, X_2, \ldots, X_p]$ have the covariance matrix $\boldsymbol{\Sigma}$ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.

Consider the linear combinations

$$\begin{aligned}
Y_1 &= \mathbf{a}_1'\boldsymbol{X} = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\
Y_2 &= \mathbf{a}_2'\boldsymbol{X} = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \\
&\vdots \\
Y_p &= \mathbf{a}_p'\boldsymbol{X} = a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p
\end{aligned}$$

Then

$$\begin{aligned}
\mathrm{Var}(Y_i) &= \mathbf{a}_i'\boldsymbol{\Sigma}\mathbf{a}_i \quad i = 1, 2, \ldots, p \\
\mathrm{Cov}(Y_i, Y_k) &= \mathbf{a}_i'\boldsymbol{\Sigma}\mathbf{a}_k \quad i, k = 1, 2, \ldots, p
\end{aligned}$$

Define

$$\begin{aligned}
\text{First principle component} &= \text{linear combination } \mathbf{a}_1'\boldsymbol{X} \text{ that maximizes} \\
&\quad \mathrm{Var}(\mathbf{a}_1'\boldsymbol{X}) \text{ subject to } \mathbf{a}_1'\mathbf{a}_1 = 1 \\
\text{Second principle component} &= \text{linear combination } \mathbf{a}_2'\boldsymbol{X} \text{ that maximizes} \\
&\quad \mathrm{Var}(\mathbf{a}_2'\boldsymbol{X}) \text{ subject to } \mathbf{a}_2'\mathbf{a}_2 = 1 \text{ and} \\
&\quad \mathrm{Cov}(\mathbf{a}_1'\boldsymbol{X}, \mathbf{a}_2'\boldsymbol{X}) = 0
\end{aligned}$$

At the ith step,

$$\begin{aligned}
i\text{th principle component} &= \text{linear combination } \mathbf{a}_i'\boldsymbol{X} \text{ that maximizes} \\
&\quad \mathrm{Var}(\mathbf{a}_i'\boldsymbol{X}) \text{ subject to } \mathbf{a}_i'\mathbf{a}_i = 1 \text{ and} \\
&\quad \mathrm{Cov}(\mathbf{a}_i'\boldsymbol{X}, \mathbf{a}_k'\boldsymbol{X}) = 0 \quad \text{for} \quad k < i
\end{aligned}$$

Results 5.1 Let $\mathbf{\Sigma}$ be the covariance matrix associated with the random vector $\boldsymbol{X}' = [X_1, X_2, \ldots, X_p]$. Let $\mathbf{\Sigma}$ have the eigenvalue-eigenvector pair $(\lambda_1, \mathbf{e}_1), (\lambda_2, \mathbf{e}_2), \ldots, (\lambda_p, \mathbf{e}_p)$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Then the ***ith principal component*** is given by

$$Y_i = \mathbf{e}_i' \boldsymbol{X} = e_{i1} X_1 + e_{i2} X_2 + \cdots + e_{ip} X_p, i = 1, 2, \ldots, p$$

With these choices,

$$\begin{aligned} \mathrm{Var}(Y_i) &= \mathbf{e}_i' \mathbf{\Sigma} \mathbf{e}_i = \lambda_i, i = 1, 2, \ldots, p \\ \mathrm{Cov}(Y_i, Y_k) &= \mathbf{e}_i' \mathbf{\Sigma} \mathbf{e}_k = 0, i \neq k \end{aligned}$$

If some λ_i are equal, the choices of corresponding coefficients vectors, $\mathbf{e}_i$, and hence Y_i are not unique.

Results 5.2 Let $\boldsymbol{X}' = [X_1, X_2, \ldots, X_p]$ have covariance matrix $\mathbf{\Sigma}$, with eigenvalue-eigenvector pairs $(\lambda_1, \mathbf{e}_1), (\lambda_2, \mathbf{e}_2), \ldots, (\lambda_p, \mathbf{e}_p)$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Let $Y_1 = \mathbf{e}_1' \boldsymbol{X}, Y_2 = \mathbf{e}_2' \boldsymbol{X}, \ldots, Y_p = \mathbf{e}_p' \boldsymbol{X}$ be the principal components. Then

$$\sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} = \sum_{i=1}^{p} \mathrm{Var}(X_i) = \lambda_1 + \lambda_2 + \cdots + \lambda_p = \sum_{i=1}^{p} \mathrm{Var}(Y_i)$$

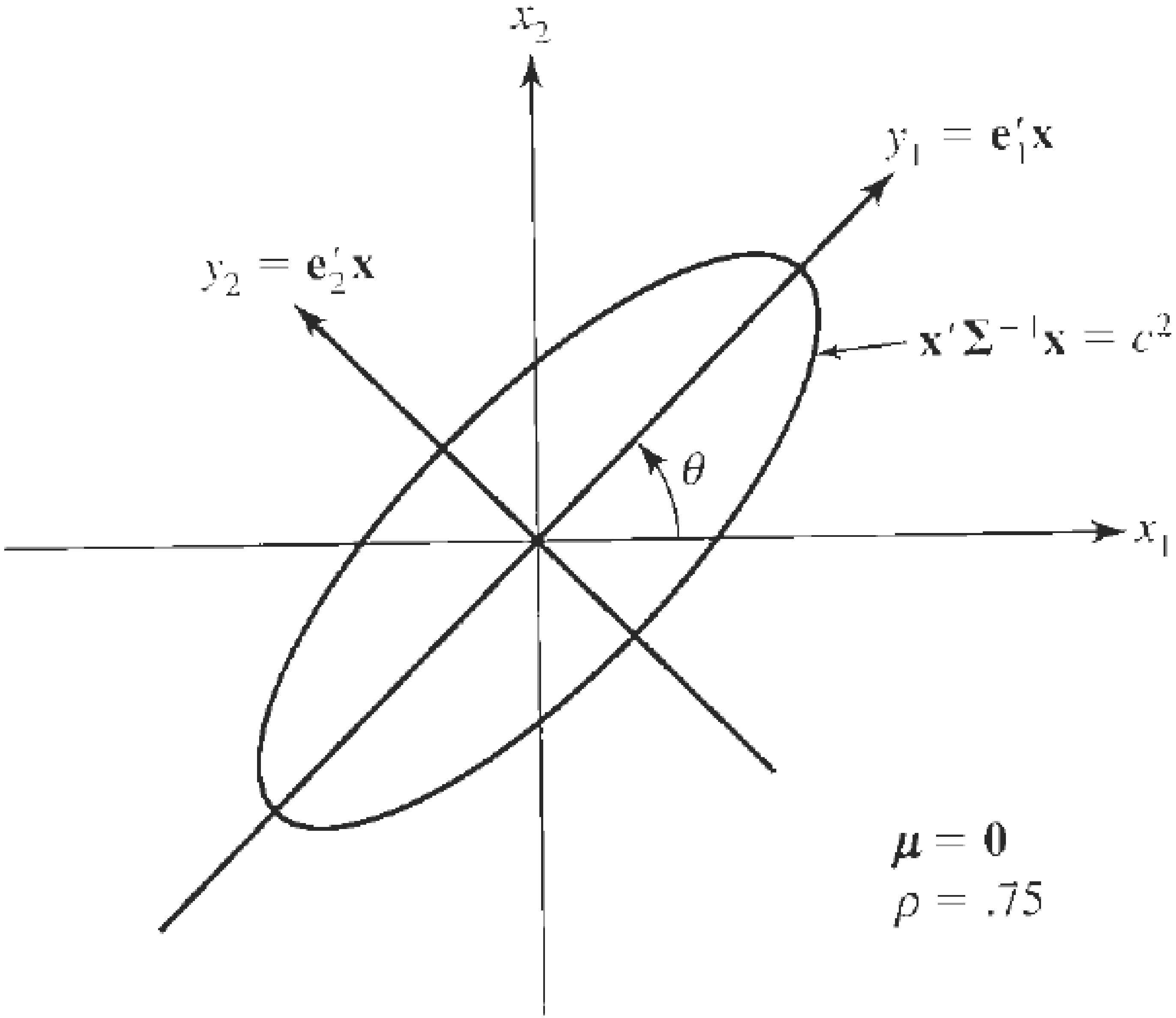


Figure 8.1 The constant density ellipse $\mathbf{x}'\mathbf{\Sigma}^{-1}\mathbf{x} = c^2$ and the principal components y_1, y_2 for a bivariate normal random vector $\mathbf{X}$ having mean $\mathbf{0}$.

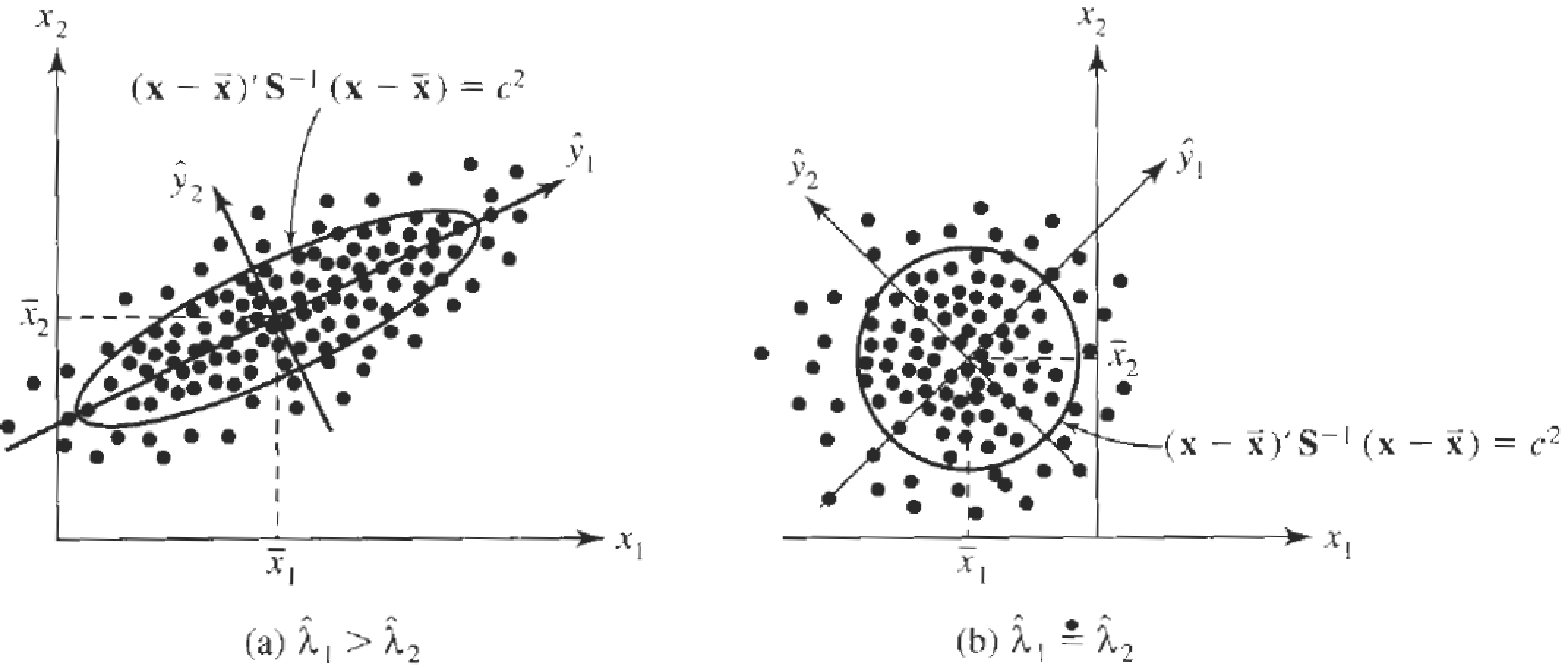


Figure 8.4 Sample principal components and ellipses of constant distance.

If A is a square matrix, a non-zero vector $\mathbf{v}$ is an eigenvector of A if there is a scalar λ (eigenvalue) such that

$$Av = \lambda v$$

Example: $\begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 12 \\ 8 \end{pmatrix} = 4 \begin{pmatrix} 3 \\ 2 \end{pmatrix}$

Matrix decomposition

Theorem 1: if square $d \times d$ matrix $\mathbf{S}$ is a real and symmetric matrix ($\mathbf{S} = \mathbf{S}^{\mathrm{T}}$) then

$$\mathbf{S} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{\mathrm{T}}$$

where $\mathbf{V} = [v_1 \quad \cdots \quad v_d]$ are the eigenvectors of $\mathbf{S}$ and $\boldsymbol{\Lambda} = diag(\lambda_1, \ldots, \lambda_d)$ are the corresponding eigenvalues.

PCA

- Find the direction for which the variance is maximized:

$$v_1 = argmax_{v1}\ var(Xv_1)$$

Subject to: $v_1{}^T v_1 = 1$

- Rewrite in terms of the covariance matrix:

$$var(Xv_1) = \frac{1}{N-1}(Xv_1)^T(Xv_1) = v_1{}^T \frac{1}{N-1} X^T X\, v_1 = v_1{}^T C\, v_1$$

- Solve via constrained optimization:

$$L(v_1, \lambda_1) = v_1{}^T C\, v_1 + \lambda_1(1 - v_1{}^T v_1)$$

- Constrained optimization:

$$L(v_1, \lambda_1) = {v_1}^T C\, v_1 + \lambda_1(1 - {v_1}^T v_1)$$

- Gradient with respect to v_1:

$$\frac{dL(v_1, \lambda_1)}{dv_1} = 2Cv_1 - 2\lambda_1 v_1 \Rightarrow Cv_1 = \lambda_1 v_1$$

This is the eigenvector problem!

- Multiply by ${v_1}^T$:

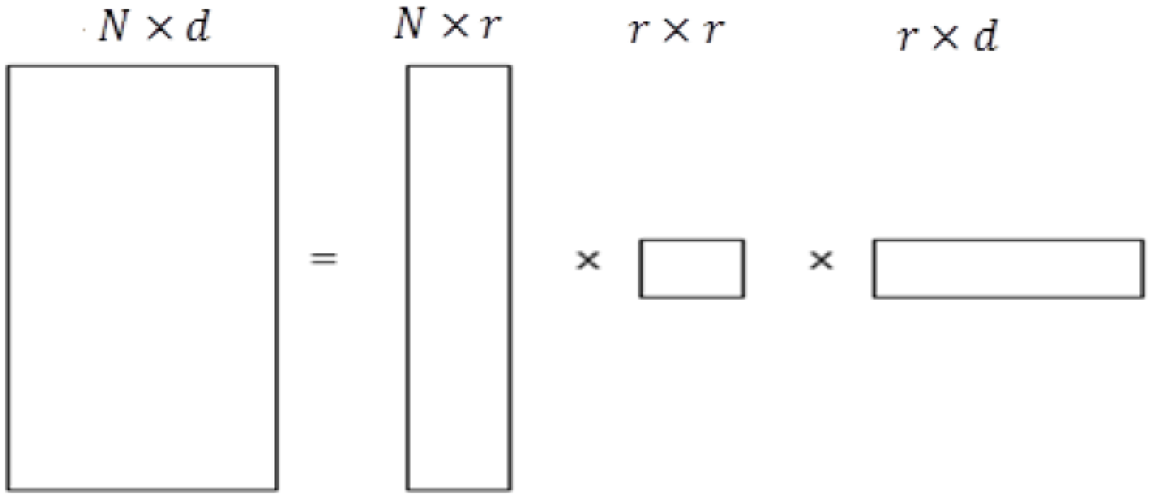
$$\lambda_1 = {v_1}^T C\, v_1$$

The projection variance is the eigenvalue

SVD

Any $N \times d$ matrix X can be uniquely expressed as:

$$\mathbf{X} = \mathbf{U} \times \mathbf{\Sigma} \times \mathbf{V}^{\mathrm{T}}$$



- r is the rank of the matrix X (# of linearly independent columns/rows).
- U is a column-orthonormal $N \times r$ matrix.
- Σ is a diagonal $r \times r$ matrix where the singular values σ_i are sorted in descending order.
- V is a column-orthonormal $d \times r$ matrix.

PCA and SVD relation

Theorem: Let $X = U\,\Sigma\,V^T$ be the SVD of an $N \times d$ matrix X and $C = \frac{1}{N-1} X^T X$ be the $d \times d$ covariance matrix. The eigenvectors of C are the same as the right singular vectors of X.

Proof:

$$X^T X = V\,\Sigma\,U^T\,U\,\Sigma\,V^T = V\,\Sigma\,\Sigma\,V^T = V\,\Sigma^2\,V^T$$

$$C = V \frac{\Sigma^2}{N-1} V^T$$

But C is symmetric, hence $C = V\,\Lambda\,V^T$ (according to theorem1).

Therefore, the eigenvectors of the covariance matrix are the same as matrix V (right singular vectors) and the eigenvalues of C can be computed from the singular values $\lambda_i = \frac{\sigma_i^2}{N-1}$

Summary for PCA and SVD

<u>Objective</u>: project an $N \times d$ data matrix X using the largest m principal components $V = [v_1, \dots v_m]$.

1. zero mean the columns of X.
2. Apply PCA or SVD to find the principle components of X.

PCA:

I. Calculate the covariance matrix $C = \frac{1}{N-1} X^T X$.

II. V corresponds to the eigenvectors of C.

SVD:

I. Calculate the SVD of $X = U \Sigma V^T$.

II. V corresponds to the right singular vectors.

3. Project the data in an m dimensional space: $Y = X V$