IAI, TCG CREST

Avisek Gupta (avisek003@gmail.com)

Machine Learning

10 – Principal Component Analysis

September 23, 2022

Preliminaries: **Symmetric Matrices**

$A \in \mathbb{R}^{d \times d}$ is a symmetric matrix iff $A^T = A$.

$If A \in \mathbb{R}^{d \times d}$ is a real and symmetric matrix, then it has real eigenvalues and orthogonal eigenvectors.

Let $A \in \mathbb{R}^{d \times d}$ be a **symmetric** matrix. Then a spectral decomposition of A can be proved to be:

$$A = \lambda_1 e_1 e_1^T + \lambda_2 e_2 e_2^T + ... + \lambda_d e_d e_d^T$$

where $\lambda_1, ..., \lambda_d$ are the eigenvalues of A, and $e_1, ..., e_d$ are the corresponding orthonormal eigenvectors ($e_i^T e_i = 1, e_i^T e_j = 0 \ \forall i, j$). Here $e_i e_i^T \in \mathbb{R}^{d \times d}$ is the outer product of the eigenvectors.

Preliminaries: **Positive Semi-Definite Matrices**

$A \in \mathbb{R}^{d \times d}$ is a positive semi-definite matrix iff $x^T A x \geq 0$.

$A \in \mathbb{R}^{d \times d}$ is a positive definite matrix iff $x^T A x > 0$.

A positive (semi-)definite matrix A has all eigenvalues greater than (or equal to) zero.

Preliminaries: **Symmetric and Positive Semi-Definite Matrices**

If $A \in \mathbb{R}^{d \times d}$ is symmetric and positive semi-definite, then we can write the spectral decomposition of A as,

$$A = \lambda_1 e_1 e_1^T + \lambda_2 e_2 e_2^T + ... + \lambda_d e_d e_d^T.$$

From the definition of positive semi-definite matrices,

$$\begin{aligned} & x^T A x \geq 0 \\ \implies & x^T A x = \lambda_1 x^T e_1 e_1^T x + ... + \lambda_d x^T e_d e_d^T x \geq 0 \\ \implies & \lambda_1 (x^T e_1)^2 + ... + \lambda_d (x^T e_d)^2 \geq 0 \\ \implies & \lambda_1 y_1^2 + ... + \lambda_d y_d^2 = c^2 \geq 0 \end{aligned}$$

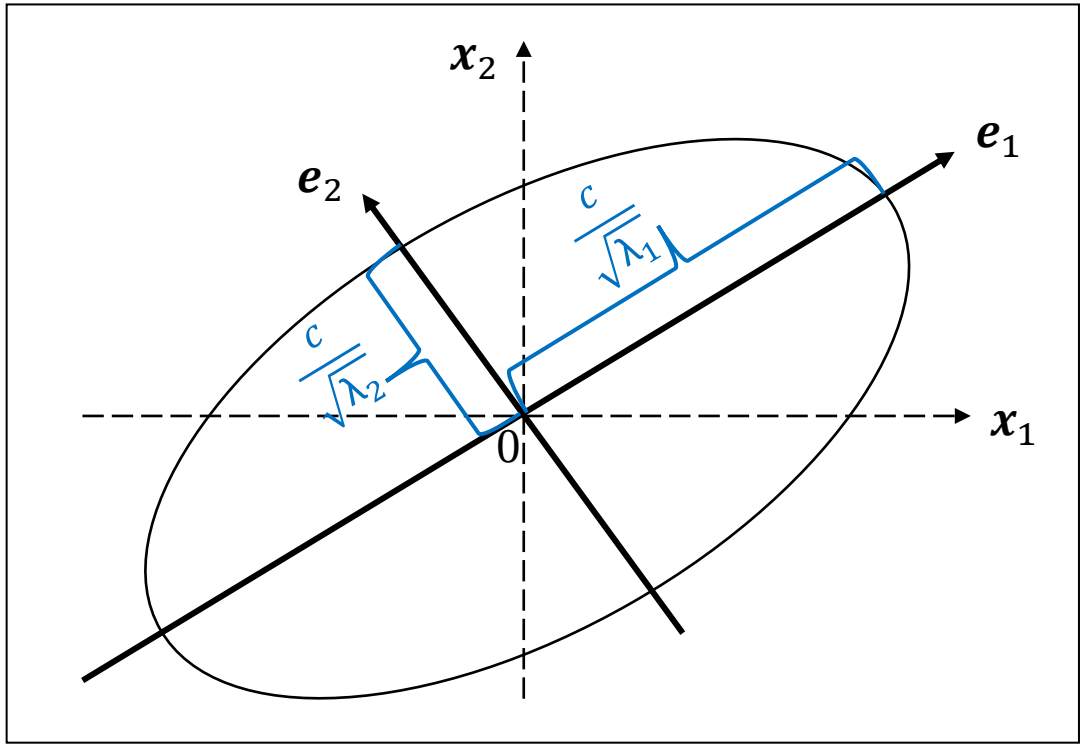
Therefore we obtain the equation of a d-dimensional ellipsoid, since $\lambda_i \geq 0 \ \forall i$.

Preliminaries: **Symmetric and Positive Semi-Definite Matrices**

For $d = 2$:

$$x^T Ax = \lambda_1 (x^T e_1)^2 + \lambda_2 (x^T e_2)^2 = c^2$$

is the loci of the points on an ellipse that have distance $\sqrt{x^T Ax}$ to the origin, with half lengths of $c/\sqrt{\lambda_i}$ in the direction of each eigenvector e_i.



Dimension Reduction

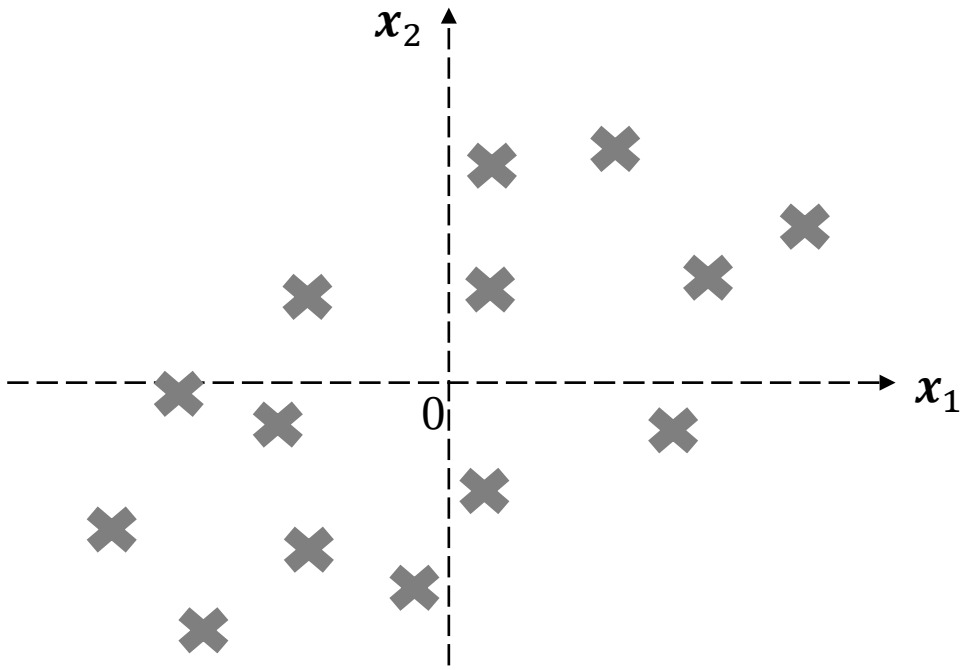
Given a dataset $X = \{\mathbf{x}_1, ..., \mathbf{x}_n\}, \mathbf{x}_i \in \mathbb{R}^d$, we wish to find a projection of X to a space of lower dimension $k < d$, in order to obtain lower-dimension representations of the data $Z = \{\mathbf{z}_1, ..., \mathbf{z}_n\}, \mathbf{z}_i \in \mathbb{R}^k$.

There are several reasons why we would want to reduce the dimensions of data:

1. **Visualize high-dimensional data**: Being able to obtain a 2-dimensional representation would be useful.
2. **The curse of dimensionality**: The volume of the space in which an n-sized sample lies increases exponentially with the increase in the number of dimensions. This makes the statistical estimation of a distribution more challenging, since the n-sized sample becomes a highly sparse representation of the distribution.
3. **The failure of distance metrics**: In high-dimensional spaces, a point ends up having approximately the same distance to all other points.

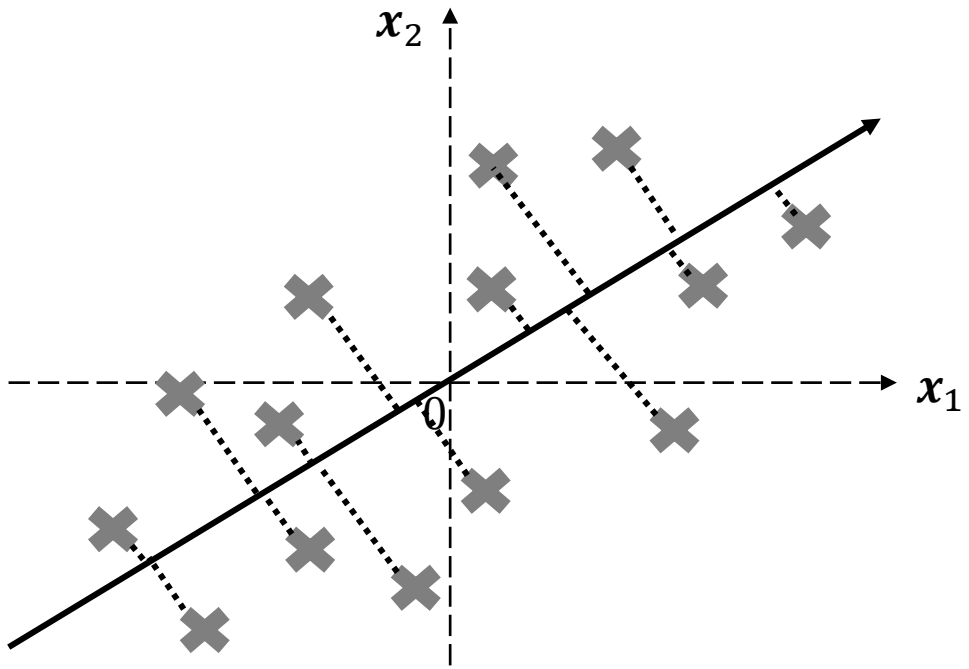
Principal Component Analysis

Given a dataset $X = [\mathbf{x}_1, ..., \mathbf{x}_n]^T, \mathbf{x}_i \in \mathbb{R}^d$, we wish to find a projection of X to a space of lower dimension $k < d$, $Z = [\mathbf{z}_1, ..., \mathbf{z}_k], \mathbf{z}_i \in \mathbb{R}^n$, such that the variance of the projections is maximized.



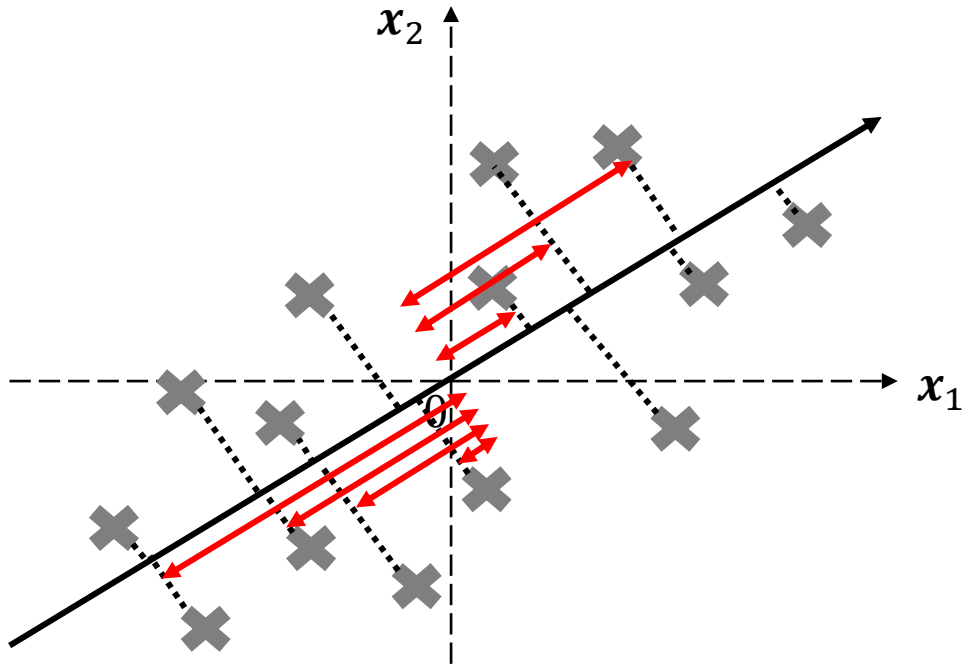
Principal Component Analysis

Given a dataset $X = [\mathbf{x}_1, ..., \mathbf{x}_n]^T, \mathbf{x}_i \in \mathbb{R}^d$, we wish to find a projection of X to a space of lower dimension $k < d$, $Z = [\mathbf{z}_1, ..., \mathbf{z}_k], \mathbf{z}_i \in \mathbb{R}^n$, such that the variance of the projections is maximized.



Principal Component Analysis

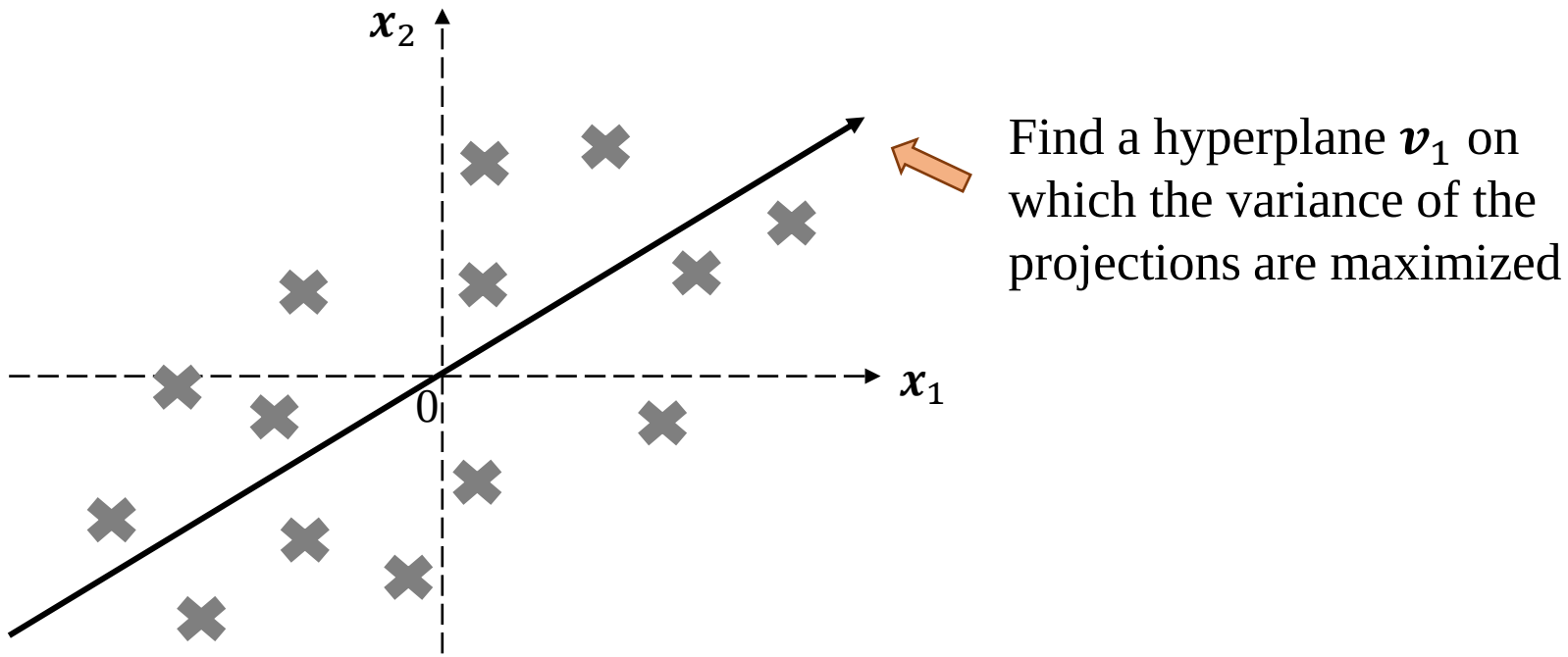
Given a dataset $X = [\mathbf{x}_1, ..., \mathbf{x}_n]^T, \mathbf{x}_i \in \mathbb{R}^d$, we wish to find a projection of X to a space of lower dimension $k < d$, $Z = [\mathbf{z}_1, ..., \mathbf{z}_k], \mathbf{z}_i \in \mathbb{R}^n$, such that the variance of the projections is maximized.



Principal Component Analysis

Given a dataset $X = [\mathbf{x}_1, ..., \mathbf{x}_n]^T, \mathbf{x}_i \in \mathbb{R}^d$, we wish to find a projection of X to a space of lower dimension $k < d$, $Z = [\mathbf{z}_1, ..., \mathbf{z}_k], \mathbf{z}_i \in \mathbb{R}^n$, such that the variance of the projections is maximized.

1. Find a hyperplane $\mathbf{v}_1$ such that for $\mathbf{z}_1 = X\mathbf{v}_1$, the variance of $\mathbf{z}_1$ is maximized, i.e., $\max_{\mathbf{v}_1} \sum_{i=1}^{n} (\mathbf{z}_{1i} - \bar{\mathbf{z}_1})^2$.



Principal Component Analysis

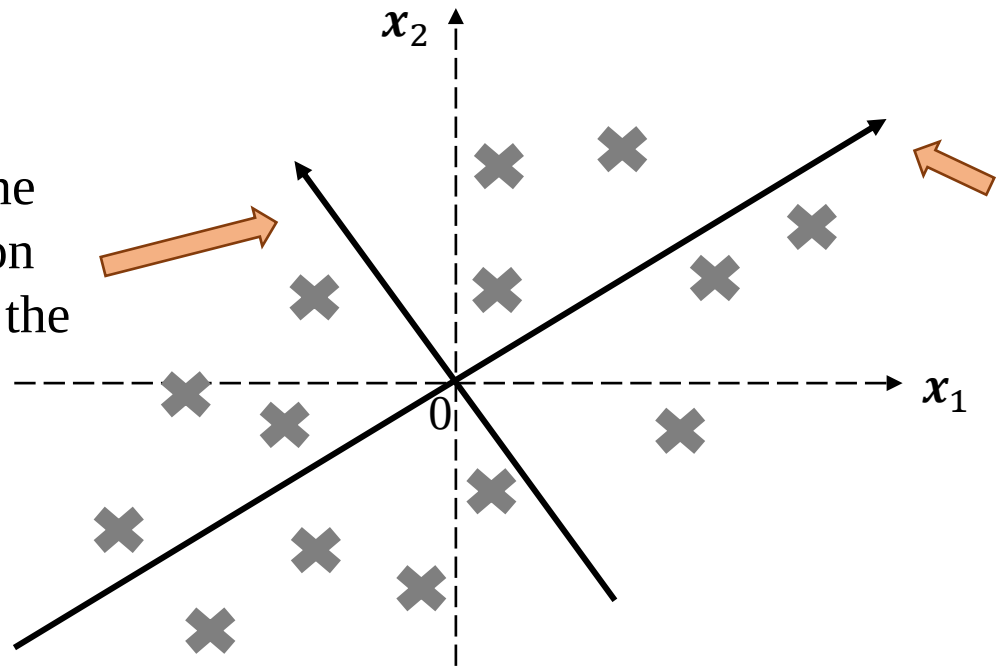
Given a dataset $X = [\mathbf{x}_1, ..., \mathbf{x}_n]^T, \mathbf{x}_i \in \mathbb{R}^d$, we wish to find a projection of X to a space of lower dimension $k < d$, $Z = [\mathbf{z}_1, ..., \mathbf{z}_k], \mathbf{z}_i \in \mathbb{R}^n$, such that the variance of the projections is maximized.

1. Find a hyperplane $\mathbf{v}_1$ such that for $\mathbf{z}_1 = X\mathbf{v}_1$, the variance of $\mathbf{z}_1$ is maximized, i.e., $\max_{\mathbf{v}_1} \sum_{i=1}^{n} (\mathbf{z}_{1i} - \bar{\mathbf{z}_1})^2$.

2. Find a hyperplane $\mathbf{v}_2$, $\mathbf{v}_2 \perp \mathbf{v}_1$ such that for $\mathbf{z}_2 = X\mathbf{v}_2$, the variance of $\mathbf{z}_2$ is maximized, i.e., $\max_{\mathbf{v}_2, \mathbf{v}_2 \perp \mathbf{v}_1} \sum_{i=1}^{n} (\mathbf{z}_{2i} - \bar{\mathbf{z}_2})^2$

Principal Component Analysis

2. Find a hyperplane $\mathbf{v}_2$, $\mathbf{v}_2 \perp \mathbf{v}_1$ such that for $\mathbf{z}_2 = X\mathbf{v}_2$, the variance of $\mathbf{z}_2$ is maximized, i.e., $\max_{\mathbf{v}_2, \mathbf{v}_2 \perp \mathbf{v}_1} \sum_{i=1}^{n} (\mathbf{z}_{2i} - \bar{\mathbf{z}}_2)^2$



Next, find a hyperplane $\boldsymbol{v}_2$ such that $\boldsymbol{v}_2 \perp \boldsymbol{v}_1$, on which the variance of the projections are maximized

Find a hyperplane $\boldsymbol{v}_1$ on which the variance of the projections are maximized

Principal Component Analysis

Given a dataset $X = [\mathbf{x}_1, ..., \mathbf{x}_n]^T, \mathbf{x}_i \in \mathbb{R}^d$, we wish to find a projection of X to a space of lower dimension $k < d$, $Z = [\mathbf{z}_1, ..., \mathbf{z}_k], \mathbf{z}_i \in \mathbb{R}^n$, such that the variance of the projections is maximized.

1. Find a hyperplane $\mathbf{v}_1$ such that for $\mathbf{z}_1 = X\mathbf{v}_1$, the variance of $\mathbf{z}_1$ is maximized, i.e., $\max_{\mathbf{v}_1} \sum_{i=1}^{n} (\mathbf{z}_{1i} - \bar{\mathbf{z}_1})^2$.

2. Find a hyperplane $\mathbf{v}_2,\ \mathbf{v}_2 \perp \mathbf{v}_1$ such that for $\mathbf{z}_2 = X\mathbf{v}_2$, the variance of $\mathbf{z}_2$ is maximized, i.e., $\max_{\mathbf{v}_2, \mathbf{v}_2 \perp \mathbf{v}_1} \sum_{i=1}^{n} (\mathbf{z}_{2i} - \bar{\mathbf{z}_2})^2$

...

k. Find a hyperplane $\mathbf{v}_k,\ s.t., \mathbf{v}_k \perp \mathbf{v}_1, \mathbf{v}_k \perp \mathbf{v}_2, ..., \mathbf{v}_k \perp \mathbf{v}_{k-1}$ such that for $\mathbf{z}_k = X\mathbf{v}_k$, the variance of $\mathbf{z}_k$ is maximized, i.e.,

$$\max_{\mathbf{v}_k, s.t., \mathbf{v}_k \perp \mathbf{v}_1, \mathbf{v}_k \perp \mathbf{v}_2, ..., \mathbf{v}_k \perp \mathbf{v}_{k-1}} \sum_{i=1}^{n} (\mathbf{z}_{ki} - \bar{\mathbf{z}_k})^2$$

Principal Component Analysis

Given a dataset $X = [\mathbf{x}_1, ..., \mathbf{x}_n]^T, \mathbf{x}_i \in \mathbb{R}^d$, we wish to find a projection of X to a space of lower dimension $k < d$, $Z = [\mathbf{z}_1, ..., \mathbf{z}_k], \mathbf{z}_i \in \mathbb{R}^n$, such that the variance of the projections is maximized.

1. Find a hyperplane $\mathbf{v}_1$ such that for $\mathbf{z}_1 = X\mathbf{v}_1$, the variance of $\mathbf{z}_1$ is maximized, i.e., $\max_{\mathbf{v}_1} \sum_{i=1}^{n} (\mathbf{z}_{1i} - \bar{\mathbf{z}_1})^2$.

...

k. Find a hyperplane $\mathbf{v}_k$, $s.t., \mathbf{v}_k \perp \mathbf{v}_1, \mathbf{v}_k \perp \mathbf{v}_2, ..., \mathbf{v}_k \perp \mathbf{v}_{k-1}$ such that for $\mathbf{z}_k = X\mathbf{v}_k$, the variance of $\mathbf{z}_k$ is maximized, i.e.,

$$\max_{\mathbf{v}_k, s.t., \mathbf{v}_k \perp \mathbf{v}_1, \mathbf{v}_k \perp \mathbf{v}_2, ..., \mathbf{v}_k \perp \mathbf{v}_{k-1}} \sum_{i=1}^{n} (\mathbf{z}_{ki} - \bar{\mathbf{z}_k})^2$$

Therefore, PCA finds $Z \in \mathbb{R}^{n \times k}$:

$$Z = [\mathbf{z}_1, ..., \mathbf{z}_k] = XV, X \in \mathbb{R}^{n \times d}, V \in \mathbb{R}^{d \times k}.$$

Solving PCA: finding $\mathbf{v}_1$

To find the *first principle component* $\mathbf{v}_1$, solve:

$$\max_{\mathbf{v}_1} \sum_{i=1}^{n} (\mathbf{z}_{1i} - \bar{\mathbf{z}_1})^2, \ s.t., \ \sum_{i=1}^{d} v_{1i}^2 = 1.$$

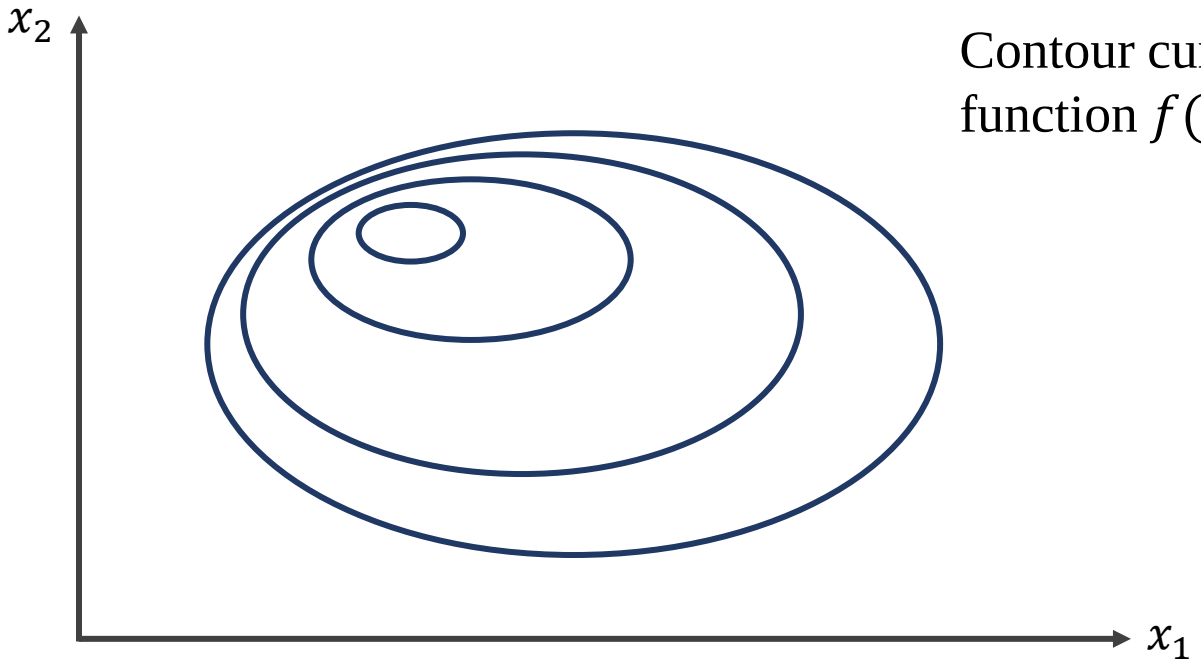
Or,

$$\max_{\mathbf{v}_1} (\mathbf{z}_1 - \bar{\mathbf{z}_1})^T (\mathbf{z}_1 - \bar{\mathbf{z}_1}), \ s.t., \ \mathbf{v}_1^T \mathbf{v}_1 = 1.$$

A deviation: Unconstrained optimization

We wish to optimize (e.g., minimize) a function $f(x_1, ..., x_n)$:

$$\min f(x_1, ..., x_n)$$



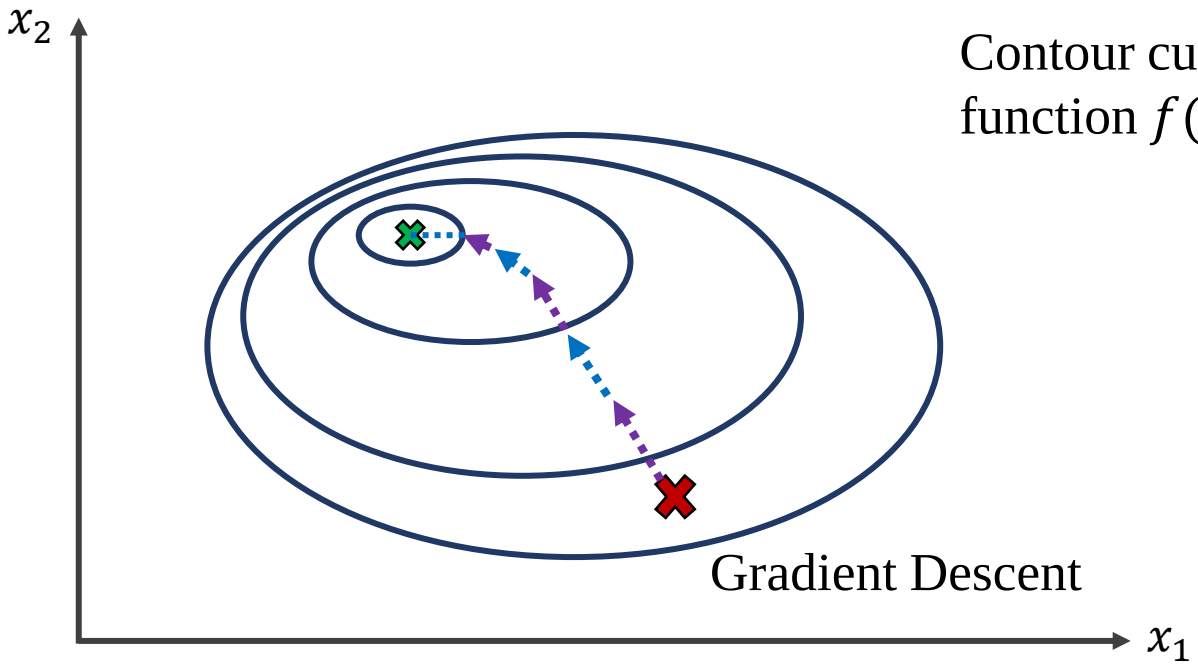
Contour curves of a function $f(x_1, x_2)$

A deviation: Unconstrained optimization

We wish to optimize (e.g., minimize) a function $f(x_1, ..., x_n)$:

$$\min f(x_1, ..., x_n)$$

Gradient Descent can be a general approach to find the optima of a **differentiable** function.



A deviation: Constrained optimization

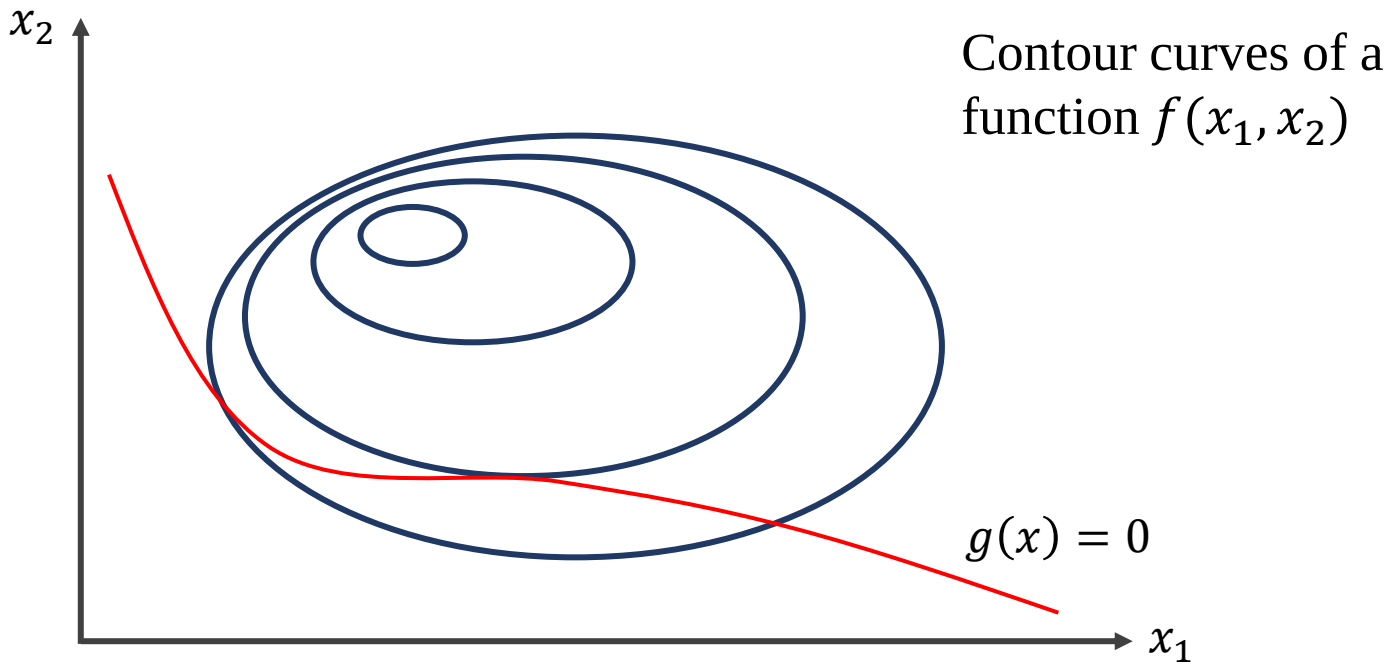
We wish to optimize (e.g., minimize) a function $f(x_1, ..., x_n)$ subject to constraints $g_1(x_1, ..., x_n) = 0, g_2(x_1, ..., x_n) = 0, ..., g_c(x_1, ..., x_n) = 0$:

$$\begin{aligned} &\min\ f(x_1, ..., x_n) \\ s.t.,\ \ &g_1(x_1, ..., x_n) = 0, \\ &g_2(x_1, ..., x_n) = 0, \\ &..., \\ &g_c(x_1, ..., x_n) = 0 \end{aligned}$$

A deviation: Constrained optimization

An example:

$$\min\ f(x_1, x_2)$$
$$s.t.,\ g(x_1, x_2) = 0$$



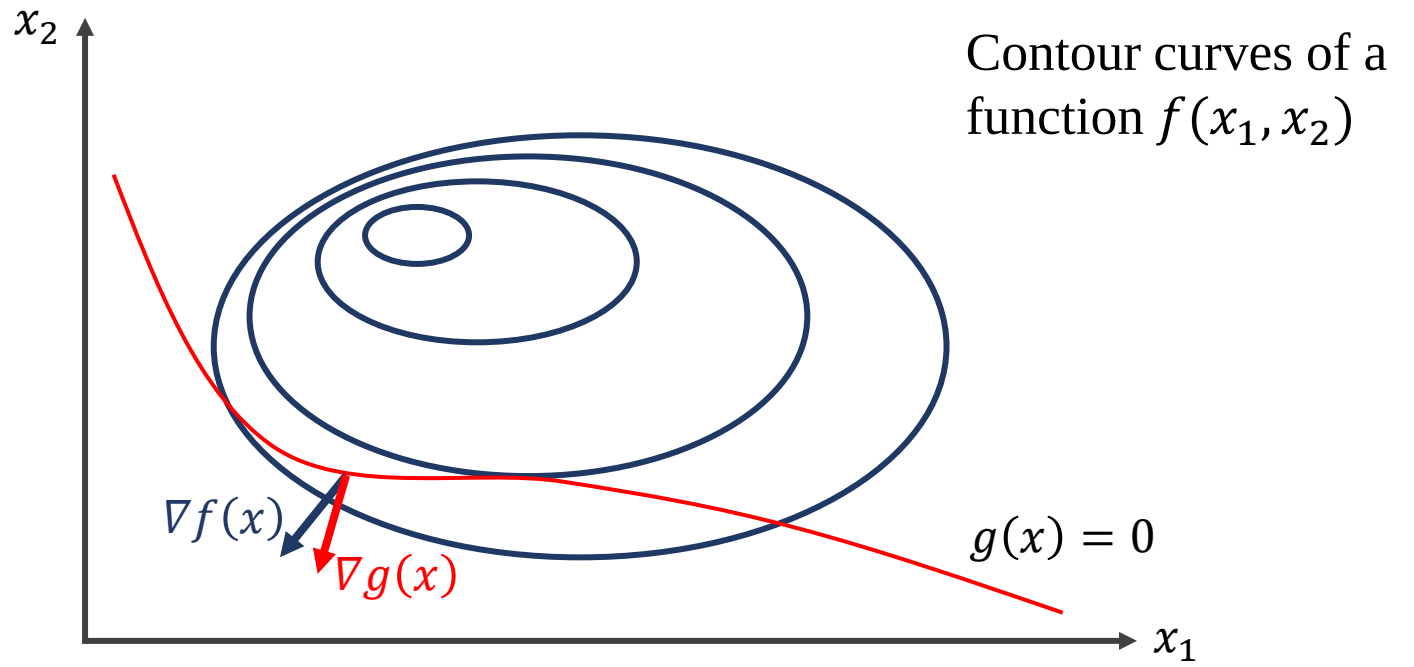
A deviation: Constrained optimization

We want to find a **local minima** at which,

$$\nabla f(x_1, x_2) = \lambda \nabla g(x_1, x_2)$$

If $\nabla f(x_1, x_2) \neq \lambda \nabla g(x_1, x_2)$, then there is a direction along $g(x_1, x_2) = 0$ in which $f(x_1, x_2)$ can be decreased.

The **lagrangian multiplier** λ is kept to make the magnitudes of the gradients equal.



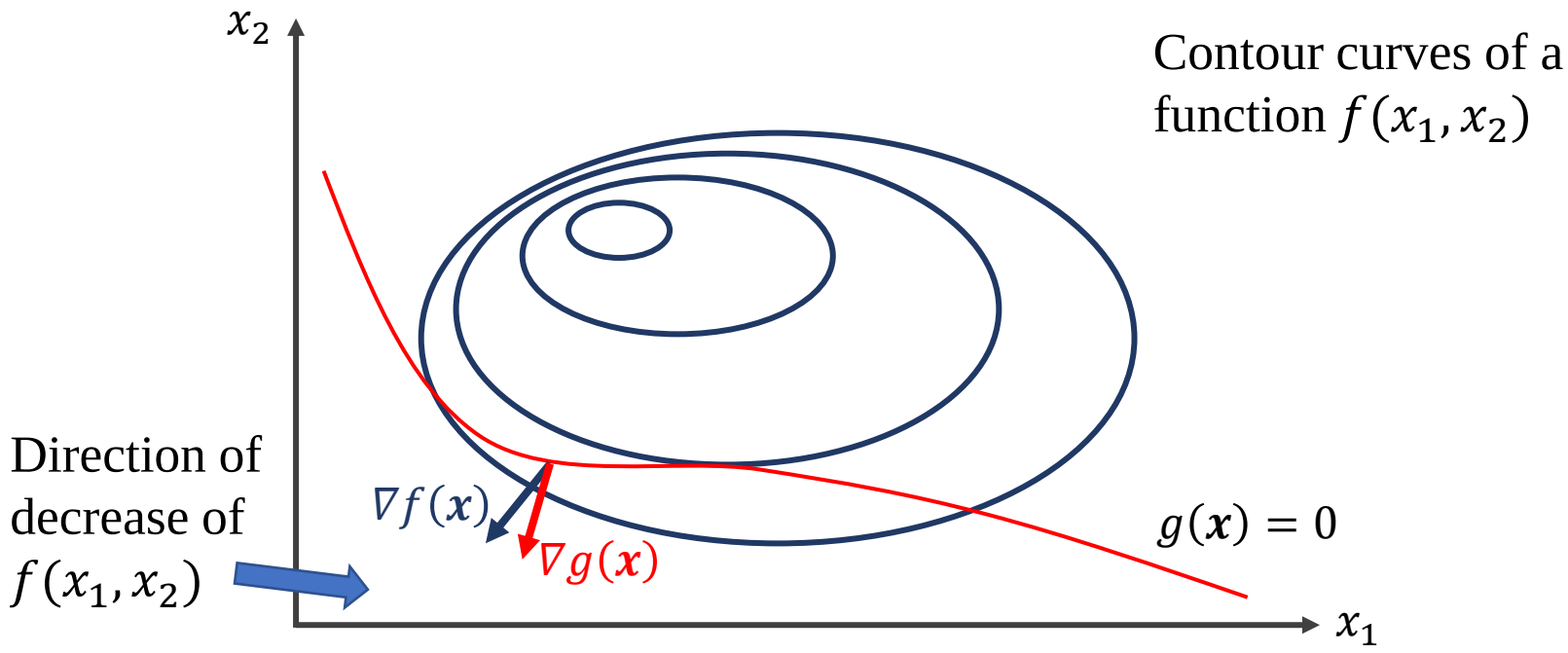
A deviation: Constrained optimization

We want to find a **local minima** at which,

$$\nabla f(x_1, x_2) = \lambda \nabla g(x_1, x_2)$$

If $\nabla f(x_1, x_2) \neq \lambda \nabla g(x_1, x_2)$, then there is a direction along $g(x_1, x_2) = 0$ in which $f(x_1, x_2)$ can be decreased.

The **lagrangian multiplier** λ is kept to make the magnitudes of the gradients equal.



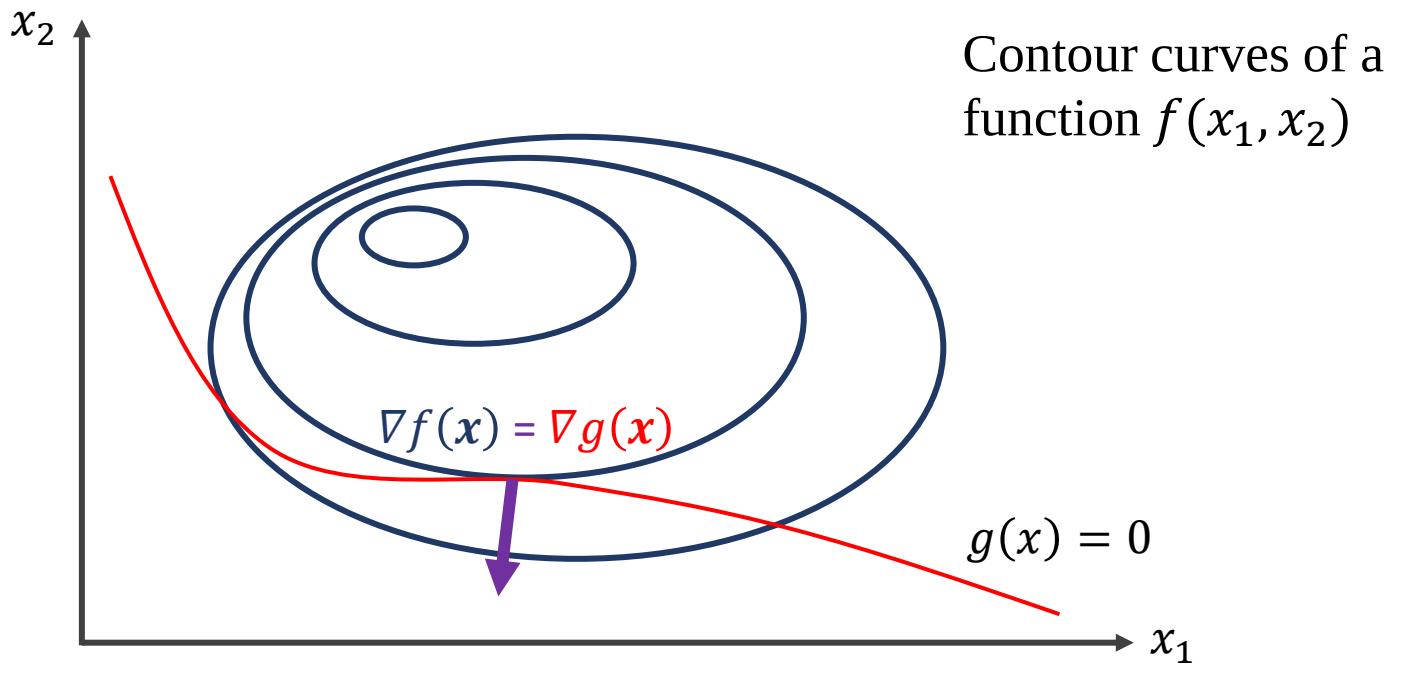
A deviation: Constrained optimization

We want to find a **local minima** at which,

$$\nabla f(x_1, x_2) = \lambda \nabla g(x_1, x_2)$$

If $\nabla f(x_1, x_2) \neq \lambda \nabla g(x_1, x_2)$, then there is a direction along $g(x_1, x_2) = 0$ in which $f(x_1, x_2)$ can be decreased.

The **lagrangian multiplier** λ is kept to make the magnitudes of the gradients equal.



A deviation: Constrained optimization

From the constrained optimization problem,

$$\begin{aligned} &\min\ f(x_1, x_2) \\ &s.t.,\ g(x_1, x_2) = 0 \end{aligned}$$

A **Lagrangian function** is defined as,

$$\mathcal{L} = f(x_1, x_2) - \lambda\, g(x_1, x_2)$$

Equating the derivatives of the Lagrangian function to zero,

$$\begin{aligned} \nabla_{\mathbf{x}}\, \mathcal{L} &= \nabla_{\mathbf{x}}\, f(x_1, x_2) - \lambda\, \nabla_{\mathbf{x}}\, g(x_1, x_2) = 0 \\ &\implies \nabla_{\mathbf{x}}\, f(x_1, x_2) = \lambda\, \nabla_{\mathbf{x}}\, g(x_1, x_2) \\ \nabla_{\lambda}\, \mathcal{L} &= \nabla_{\lambda}\, \lambda\, g(x_1, x_2) = 0 \\ &\implies g(x_1, x_2) = 0 \end{aligned}$$

A deviation: Constrained optimization

In general for a constrained optimization problem of the form:

$$\min\ f(\mathbf{x})$$
$$s.t.,\ g_1(\mathbf{x}) = 0,\ \ ...,\ \ g_c(\mathbf{x}) = 0$$

We define a Lagrangian function:

$$\min\ \mathcal{L} = f(\mathbf{x}) - \lambda_1\, g_1(\mathbf{x}) - ... - \lambda_c\, g_c(\mathbf{x})$$

Equating the derivatives of the Lagrangian function to zero,

$$\nabla_{\mathbf{x}}\, \mathcal{L} = \nabla_{\mathbf{x}}\, f(\mathbf{x}) - \sum_{i=1}^{c} \lambda_i\, \nabla_{\mathbf{x}}\, g_i(\mathbf{x}) = 0$$

$$\implies\ \nabla_{\mathbf{x}}\, f(\mathbf{x}) = \sum_{i=1}^{c} \lambda_i\, \nabla_{\mathbf{x}}\, g_i(\mathbf{x})$$

$$\nabla_{\lambda_i}\, \mathcal{L} = \nabla_{\lambda_i}\, \lambda_i\, g_i(\mathbf{x}) = 0$$

$$\implies\ g_i(\mathbf{x}) = 0$$

Solving PCA: finding $\mathbf{v}_1$

To find the *first principle component* $\mathbf{v}_1$, solve:

$$\max_{\mathbf{v}_1} (\mathbf{z}_1 - \bar{\mathbf{z}_1})^T(\mathbf{z}_1 - \bar{\mathbf{z}_1}),\ s.t.,\ \mathbf{v}_1^T\mathbf{v}_1 = 1.$$

Forming the Lagrangian,

$$\max_{\mathbf{v}_1} \mathcal{L} = (\mathbf{z}_1 - \bar{\mathbf{z}_1})^T(\mathbf{z}_1 - \bar{\mathbf{z}_1}) - \lambda(\mathbf{v}_1^T\mathbf{v}_1 - 1)$$

Let $\bar{X} = 0$. Then, $\bar{\mathbf{z}_1} = \frac{1}{n}\sum_{i=1}^n z_{1i} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{v}_1 = \mathbf{v}_1\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i = 0$

Then the Lagrangian is,

$$\begin{aligned}\max_{\mathbf{v}_1} \mathcal{L} &= \mathbf{z}_1^T\mathbf{z}_1 - \lambda(\mathbf{v}_1^T\mathbf{v}_1 - 1) \\ &= (X\mathbf{v}_1)^T(X\mathbf{v}_1) - \lambda(\mathbf{v}_1^T\mathbf{v}_1 - 1) \\ &= \mathbf{v}_1^T X^T X\mathbf{v}_1 - \lambda(\mathbf{v}_1^T\mathbf{v}_1 - 1) \\ &= \mathbf{v}_1^T C\mathbf{v}_1 - \lambda(\mathbf{v}_1^T\mathbf{v}_1 - 1)\end{aligned}$$

where $C \in \mathbb{R}^{d\times d}$ is the covariance matrix of $X \in \mathbb{R}^{n\times d}$.

Solving PCA: finding $\mathbf{v}_1$

To find $\mathbf{v}_1$, solve (with $\bar{X} = 0$):

$$\max_{\mathbf{v}_1} \ \mathcal{L} = \mathbf{v}_1^T C \mathbf{v}_1 - \lambda(\mathbf{v}_1^T \mathbf{v}_1 - 1)$$

Equating the derivative of $\mathcal{L}$ to zero,

$$\begin{aligned} \nabla \mathcal{L} &= C\mathbf{v}_1 - \lambda \mathbf{v}_1 = 0 \\ \implies & C\mathbf{v}_1 = \lambda \mathbf{v}_1 \end{aligned}$$

Thus $\mathbf{v}_1$ is a normalized eigenvector of C. (Which eigenvector?)

$$\max \ \mathcal{L} = \mathbf{v}_1^T C \mathbf{v}_1 - \lambda(\mathbf{v}_1^T \mathbf{v}_1 - 1) = \mathbf{v}_1^T \lambda_1 \mathbf{v}_1 = \lambda_1$$

$\mathbf{v}_1$ is the eigenvector corresponding to the **largest eigenvalue** (λ_1) of C.

Solving PCA: finding $\mathbf{v}_1, \mathbf{v}_2$

The solution of

$$\max_{\mathbf{v}_1} \mathcal{L} = \mathbf{v}_1^T C \mathbf{v}_1 - \lambda(\mathbf{v}_1^T \mathbf{v}_1 - 1)$$

is the eigenvector $\mathbf{v}_1$ corresponding to the **largest eigenvalue** (λ_1) of C.

To find the second principal component, solve:

$$\max_{\mathbf{v}_2, \mathbf{v}_2 \perp \mathbf{v}_1} \mathcal{L} = \mathbf{v}_2^T C \mathbf{v}_2 - \lambda(\mathbf{v}_2^T \mathbf{v}_2 - 1)$$

$\mathbf{v}_2$ is the eigenvector corresponding to the **second largest eigenvalue** (λ_2) of C.

Solving PCA: finding $\mathbf{v}_k$

The solution of

$$\max_{\mathbf{v}_k, s.t., \mathbf{v}_k \perp \mathbf{v}_1, \mathbf{v}_k \perp \mathbf{v}_2, ..., \mathbf{v}_k \perp \mathbf{v}_{k-1}} \mathcal{L} = \mathbf{v}_k^T C \mathbf{k}_2 - \lambda(\mathbf{v}_k^T \mathbf{v}_k - 1)$$

$\mathbf{v}_k$ is the eigenvector corresponding to the **k-th largest eigenvalue** (λ_k) of C.

The lower-dimensional PCA projection of $X \in \mathbb{R}^{n \times d}$ is $Z \in \mathbb{R}^{n \times k}$, given by,

$$Z = [\mathbf{z}_1, ..., \mathbf{z}_k] = X[\mathbf{v}_1, ..., \mathbf{v}_k] = XV, \ V \in \mathbb{R}^{d \times k}.$$

where $\mathbf{v}_i$ is the eigenvector corresponding to the i-th largest eigenvalue of the covariance of X.