Chapter 10. Dimensionality Reduction with PCA

Chapter 10. Dimensionality Reduction with PCA

10.1 Problem Setting

Consider a data set $\mathcal{X} = \{x_1, \ldots, x_N\}, x_n \in \mathbb{R}^D$ with mean 0 and the data covariance matrix:

$$S = \frac{1}{N} \sum_{n=1}^{N} x_n x_n^T. \tag{1}$$

Assume that there exists a low-dimensional compressed representation (code) of x_n

$$z_n = B^T x_n \in \mathbb{R}^M, \tag{2}$$

where

$$B := \left[b_1 \cdots b_M\right] \in \mathbb{R}^{D \times M} \text{ the projection matrix is orthonormal.} \tag{3}$$

We seek an M-dimensional subspace U of $\mathbb{R}^D$ onto which we project the data. Denote the projected data by $\tilde{x}_n \in U$, and their coordinates by z_n.

Our aim is to find projections $\tilde{x}_n$ so that they are as similar to the original data x_n and minimize the loss due to compression.

Direction with Maximal Variance:

Seek a single vector $b_1 \in \mathbb{R}^D$ that maximizes the variance of the projected data, that is we aim to maximize the variance of the first coordinate z_1 of $z \in \mathbb{R}^M$ so that

$$V_1 := V[z_1] = \frac{1}{N}\sum_{n=1}^{N} z_{1n}^2 = \frac{1}{N}\sum_{n=1}^{N}(b_1^T x_n)^2 = \frac{1}{N}\sum_{n=1}^{N} b_1^T x_n x_n^T b_1 \tag{4}$$

$$= b_1^T \left(\frac{1}{N}\sum_{n=1}^{N} x_n x_n^T\right) b_1 = b_1^T S b_1 .$$

We lead to a constrained optimization problem

$$\max_{b_1} b_1^T S b_1 \tag{5}$$

$$\text{subject to } \|b_1\|^2 = 1 \tag{6}$$

which has the Lagrangian

$$\mathcal{L}(b_1, \lambda) = b_1^T S b_1 - \lambda_1 (1 - b_1^T b_1) \tag{7}$$

Setting the partial derivatives of $\mathcal{L}$ to 0

$$\frac{\partial \mathcal{L}}{\partial b_1} = 2b_1^T S - 2\lambda_1 b_1^T = 0, \tag{8}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda_1} = 1 - b_1^T b_1 = 0, \tag{9}$$

we obtain

$$Sb_1 = \lambda_1 b_1, \ \ b_1^T b_1 = 1. \tag{10}$$

Hence b_1 is an eigenvector corresponding to the eigenvalue λ_1 of the data covariance matrix S, and

$$V_1 = b_1^T S b_1 = \lambda_1. \tag{11}$$

λ_1 is called the first principal component.

M-dimensional Subspace

Assume that we have found the first $m-1$ principal components $b_1, \ldots, b_{m-1}$ as the $m-1$ eigenvectors of S that are associated with the largest $m-1$ eigenvalues $\lambda_1, \ldots, \lambda_{m-1}$.

Since S is symmetric, these eigenvectors form an orthonormal basis of an $(m-1)$-dimensional subspace of $\mathbb{R}^D$. We find principal components that compress the remaining information.

Set $X := \begin{bmatrix} x_1 \ldots x_N \end{bmatrix} \in \mathbb{R}^{D \times N}$ and $B_{m-1} := \sum_{i=1}^{m-1} b_i b_i^T$ the projection matrix that projects onto the subspace spanned by $b_1, \ldots, b_{m-1}$. The new data matrix

$$\hat{X} := X - \sum_{i=1}^{m-1} b_i b_i^T X = X - B_{m-1} X. \tag{12}$$

M-dimensional Subspace

Let $\hat{S}$ be the data covariance matrix of the transformed dataset $\hat{\mathcal{X}} := \{\hat{x}_1, \ldots, \hat{x}_N\}$.

To find the mth principal component, we solve a constrained optimization problem

$$\max_{b_m} \left[V_m = V_m[z_m] = \frac{1}{N} \sum_{n=1}^{N} z_{mn}^2 = \frac{1}{N} \sum_{n=1}^{N} (b_m^T \hat{x}_n)^2 = b_m^T \hat{S} b_m \right] \tag{13}$$

$$\text{subject to } \|b_m\|^2 = 1.$$

The optimal solution b_m is the eigenvector of $\hat{S}$ that is associated with the largest eigenvalue of $\hat{S}$.

M-dimensional Subspace

We shall show that the sets of eigenvectors of S and $\hat{S}$ are identical.

Consider and eigenvector b_i of S, i.e., $Sb_i = \lambda_i b_i$. We have

$$\begin{aligned}\hat{S}b_i &= \frac{1}{N}\hat{X}\hat{X}^T b_i = \frac{1}{N}(X - B_{m-1}X)(X - B_{m-1}X)^T b_i \\ &= (S - SB_{m-1} - B_{m-1}S + B_{m-1}SB_{m-1})b_i. \end{aligned} \tag{14}$$

If $i \geq m$, then b_i is orthogonal to the first $m-1$ principal components and $B_{m-1}b_i = 0$. Hence

$$\hat{S}b_i = (S - B_{m-1}S)b_i = Sb_i = \lambda_i b_i, \tag{15}$$

which shows that b_m is an eigenvector of $\hat{S}$ and λ_m is the largest eigenvalue of $\hat{S}$ and λ_m is the mth largest eigenvalue of S, and both have the associated eigenvector b_m.

M-dimensional Subspace

If $i < m$, then b_i is a basis vector of the principal subspace onto which B_{m-1} projects. Since $b_1, \ldots, b_{m-1}$ are an ONB of this principal subspace, we obtain $B_{m-1} b_i = b_i$. Thus,

$$\hat{S} b_i = (S - S B_{m-1} - B_{m-1} S + B_{m-1} S B_{m-1}) b_i = 0 = 0 b_i. \tag{16}$$

This means that $b_1, \ldots, b_{m-1}$ are also eigenvectors of $\hat{S}$, but they are associated with eigenvalue 0 so that $b_1, \ldots, b_{m-1}$ span the null space of $\hat{S}$.

M-dimensional Subspace

By (15), the variance of the data projected onto the mth principal component is

$$V_m = b_m^T S b_m = \lambda_m b_m^T b_m = \lambda_m. \tag{17}$$

To find an M-dimensional subspace of $\mathbb{R}^D$ that retains as much information as possible, PCA tells us to choose the columns of the matrix B in (3) as the M eigenvectors of the data covariance matrix S that are associated with the M largest eigenvalues. The maximum amount of variance PCA can capture with the first M principal components is

$$V_M = \sum_{m=1}^{M} \lambda_m \tag{18}$$

where the λ_m are the M largest eigenvalues of the data covariance matrix S. Consequently, the variance lost by data compression via PCA is

$$J_M := \sum_{j=M+1}^{D} \lambda_i = V_D - V_M. \tag{19}$$

Setting and Objective We will derive PCA as an algorithm that directly minimizes the average reconstruction error.

Assume that $B = \{b_1, \ldots, b_D\}$ is an orthonormal basis of $\mathbb{R}^D$. Any $x \in \mathbb{R}^D$ can be written as

$$x = \sum_{d=1}^{D} \xi_d b_d = \sum_{m=1}^{M} \xi_m b_m + \sum_{j=M+1}^{D} \xi_j b_j \text{ where } \xi_d \in \mathbb{R}. \tag{20}$$

Find a vectors $\tilde{x}$ in an M-dimensional subspace U of $\mathbb{R}^D$ so that

$$\tilde{x} = \sum_{m=1}^{M} z_m b_m \tag{21}$$

is as similar to x as possible.

Without loss of generality, assume that the dataset $\mathcal{X} = \{x_1, \ldots, x_N\}$ is centered at 0, $E[\mathcal{X}] = 0$. We interested in finding the best linear projection of $\mathcal{X}$ onto U the principal subspace and ONB $b_1, \ldots, b_M$ of U. The projections of the data points

$$\tilde{x}_n := \sum_{m=1}^{M} z_{mn} b_m = B z_n \in \mathbb{R}^D. \tag{22}$$

The reconstruction error

$$J_M := \frac{1}{N} \sum_{n=1}^{N} \|x_n - \tilde{x}_n\|^2. \tag{23}$$

To find the coordinates $z_n = [z_{1n}, \ldots, z_{Mn}]^T \in \mathbb{R}^M$ of $\tilde{x}_n$ w.r.t the basis $(b_1, \ldots, b_M)$ and the ONB of the principal subspace, we follow two-step. First, we optimize the coordinate z_n for a given ONB $(b_1, \ldots, b_M)$; second, we find the optimal ONB.

Finding Optimal Coordinates

We have

$$\frac{\partial J_M}{\partial z_{in}} = \frac{\partial J_M}{\partial \tilde{x}_n}\frac{\partial \tilde{x}_n}{\partial z_{in}},$$
$$\frac{\partial J_M}{\partial \tilde{x}_n} = -\frac{2}{N}(x_n - \tilde{x}_n)^T \in \mathbb{R}^{1\times D} \tag{24}$$
$$\frac{\partial \tilde{x}_n}{\partial z_{in}} = \frac{\partial}{\partial z_{in}}\left(\sum_{m=1}^{M} z_{mn} b_m\right) = b_i.$$

Hence

$$\frac{\partial J_M}{\partial z_{in}} = -\frac{2}{N}(x_n - \tilde{x}_n)^T b_i = -\frac{2}{N}\left(x_n - \sum_{m=1}^{M} z_{mn} b_m\right)^T b_i \tag{25}$$
$$= -\frac{2}{N}(x_n^T b_i - z_{in} b_i^T b_i) = -\frac{2}{N}(x_n^T b_i - z_{in}). \tag{26}$$

Setting this to 0 obtain

$$z_{in} = x_n^T b_i = b_i^T x_n. \tag{27}$$

Finding the Basis of the Principal Subspace

To determine the basis vectors $b_1, \ldots, b_M$ of the principal subspace, we rephrase the loss function (23) using the results we have so far. To reformulate the loss function, we exploit our results from before

$$\tilde{x}_n = \sum_{m=1}^{M} z_{mn} b_m = \sum_{m=1}^{M} (x_n^T b_m) b_m = \left(\sum_{m=1}^{M} b_m b_m^T \right) x_n. \tag{28}$$

$$x_n = \sum_{d=1}^{D} z_{dn} b_d \overset{(27)}{=} \sum_{d=1}^{D} (x_n^T b_d) b_d = \left(\sum_{d=1}^{D} b_d b_d^T \right) x_n$$

$$= \left(\sum_{m=1}^{M} b_m b_m^T \right) x_n + \left(\sum_{j=M+1}^{D} b_j b_j^T \right) x_n. \tag{29}$$

Hence the difference vector between the original data point and its projection is

$$x_n - \tilde{x}_n = \left(\sum_{j=M+1}^{D} b_j b_j^T \right) x_n = \sum_{j=M+1}^{D} (x_n^T b_j) b_j. \tag{30}$$

Reformulate the loss function

$$J_M = \frac{1}{N}\sum_{n=1}^{N}\|x_n - \tilde{x}_n\|^2 \overset{(30)}{=} \frac{1}{N}\sum_{n=1}^{N}\left\|\sum_{j=M+1}^{D}(b_j^T x_n)b_j\right\|^2 \tag{31}$$

$$= \frac{1}{N}\sum_{n=1}^{N}\sum_{j=M+1}^{D}(b_j^T x_n)^2 = \frac{1}{N}\sum_{n=1}^{N}\sum_{j=M+1}^{D} b_j^T x_n b_j^T x_n \tag{32}$$

$$= \frac{1}{N}\sum_{n=1}^{N}\sum_{j=M+1}^{D} b_j^T x_n x_n^T b_j = \sum_{j=M+1}^{D} b_j^T \left(\frac{1}{N}\sum_{n=1}^{N} x_n x_n^T\right) b_j = \sum_{j=M+1}^{D} b_j^T S b_j \tag{33}$$

$$= \sum_{j=M+1}^{D} tr(b_j^T S b_j) = \sum_{j=M+1}^{D} tr(S b_j b_j^T) = tr\left(\left(\sum_{j=M+1}^{D} b_j b_j^T\right) S\right). \tag{34}$$

Hence the followings are equivalent

- Minimize the average squared reconstruction error.
- Minimize the variance of the data when projected onto the orthogonal complement of the principal subspace.
- Maximize the variance of the projection that we retain in the principal subspace which links the projection loss immediately to the maximum-variance formulation of PCA.

The average squared reconstruction error when projecting onto the M-dimensional principal subspace is

$$J_M = \sum_{j=M+1}^{D} \lambda_j, \tag{35}$$

where λ_i are the eigenvalues of the data covariance matrix. Therefore, to minimize (35) we need to select the smallest $D - M$ eigenvalues, which then implies that their corresponding eigenvectors are the basis of the orthogonal complement of the principal subspace. Consequently, the basis of the principal subspace comprises the eigenvectors $b_1, \ldots, b_M$ that are associated with the largest M eigenvalues of the data covariance matrix.

10.4 Eigenvector Computation and Low-Rank Approximations

The SVD of data matrix $X = \begin{bmatrix} x_1 & \cdots & x_N \end{bmatrix} \in \mathbb{R}^{D \times N}$ is given by

$$X = U \Sigma V^T, \tag{36}$$

where $U \in \mathbb{R}^{D \times D}$, $V \in \mathbb{R}^{N \times N}$ are orthogonal matrices and $\Sigma \in \mathbb{R}^{D \times N}$ is a matrix whose only nonzero entries are singular values $\sigma_{ii} \geq 0$. The data covariance matrix

$$S = \frac{1}{N} X X^T = \frac{1}{N} U \Sigma V^T V \Sigma^T U^T = \frac{1}{N} U \Sigma \Sigma^T U^T \tag{37}$$

The columns of U are the eigenvectors of XX^T (see Section 4.5). Futhermore, the eigenvalues λ_d of S are related to the singular values of X via

$$\lambda_d = \frac{\sigma_d^2}{N} \tag{38}$$

This relation provides the connection between the maximum variance view and the singular value decomposition.

PCA Using Low-Rank Matrix Approximations

To maximize the variance of the projected data, PCA chooses the columns of U in (37) to be the eigenvectors that are associated with the M largest eigenvalues of the data covariance matrix S so that we identify U as the projection matrix B in (3), which projects the original data onto a lower-dimensional subspace of dimension M. The Eckart-Young theorem (Section 4.6) offers a direct way to estimate the low-dimensional representation. Consider the best rank-M approximation of X

$$\tilde{X}_M := \arg \min_{rk(A) \leq M} \|X - A\|_2 \in \mathbb{R}^{D \times N}, \tag{39}$$

where $\|\cdot\|_2$ is the spectral norm defined in Section 4.6.

Practical Aspects

For large size matrices, this is not possible to find all eigenvalues and eigenvectors. In practice, we solve for eigenvalues or singular values using iterative methods, which are implemented in all modern packages for linear algebra.

In PCA, we only require a few eigenvectors. If we are interested in only the first few eigenvectors (with the largest eigenvalues), then iterative processes, which directly optimize these eigenvectors, are computationally more efficient than a full eigendecomposition (or SVD). In the extreme case of only needing the first eigenvector, a simple method called the power iteration: chooses a random vector x_0 not in the null space of S and follows the iteration

$$x_{k+1} = \frac{Sx_k}{\|Sx_k\|} \tag{40}$$

This sequence of vectors converges to the eigenvector associated with the largest eigenvalue of S.

We will provide a method for PCA when fewer data points than dimensions $N << D$.

Assume we have a centered dataset $x_1, \ldots, x_N \in \mathbb{R}^D$ and there are no duplicate data points. So $rk(S) = N$, S has $D - N + 1$ many eigenvalues that are 0. We will exploit this and turn the $D \times D$ covariance matrix into an $N \times N$ covariance matrix whose eigenvalues are all positive.

Let $(b_m, m = 1, \ldots, M)$ is a basis vector of the principal subspace.

$$Sb_m = \frac{1}{N} XX^T b_m = \lambda_m b_m \tag{41}$$

$$\Rightarrow \frac{1}{N} X^T XX^T b_m = \lambda_m X^T b_m \Leftrightarrow \frac{1}{N} X^T X c_m = \lambda_m c_m. \tag{42}$$

The eigenvector of the $N \times N$ matrix $\frac{1}{N} X^T X \in \mathbb{R}^{N \times N}$ associated with λ_m as $c_m := X^T b_m$ ($\frac{1}{N} X^T X$ and S have the same nonzero eigenvalues).

Recover the original eigenvectors via the eigenvectors of $\frac{1}{N} X^T X$

$$\frac{1}{N} X X^T X c_m = \lambda_m X c_m, \text{ which means } X c_m \text{ as an eigenvector of } S. \quad (43)$$

10.6 Key Steps of PCA in Practice

1. Mean subtraction
2. Standardization
3. Eigendecomposition of the covariance matrix
4. Projection

Summary

We have studied:

- Maximum variance perspective of PCA;
- PCA as an algorithm that directly minimizes the average reconstruction error;
- PCA using low-rank matrix approximations;
- PCA in high dimensions.