Inclass 20: Principal Component Analysis

[SCS4049] Machine Learning and Data Science

Seongsik Park (s.park@dgu.edu)

Department of Artificial Intelligence, Dongguk University

Dimensional reduction

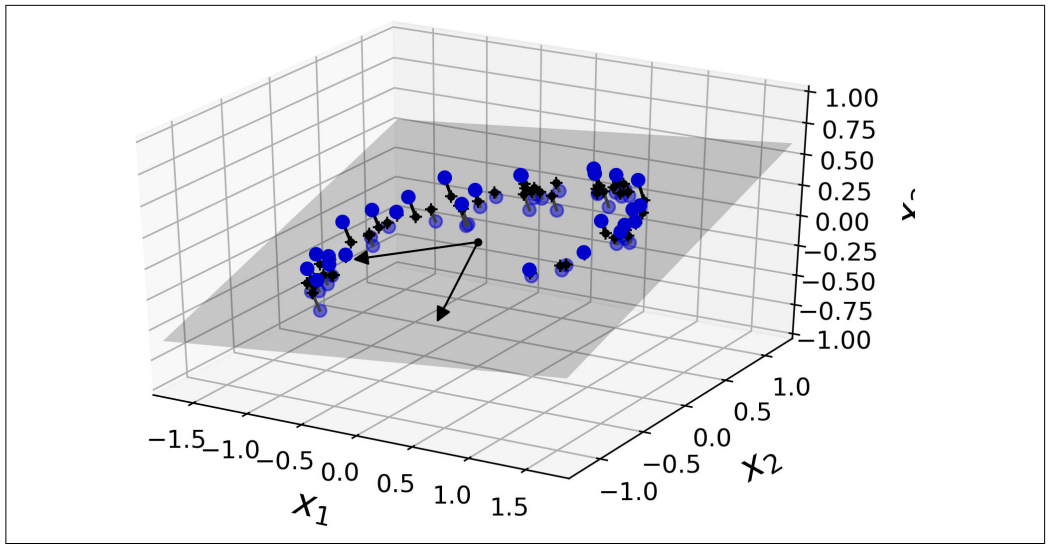


Figure 8-2. A 3D dataset lying close to a 2D subspace

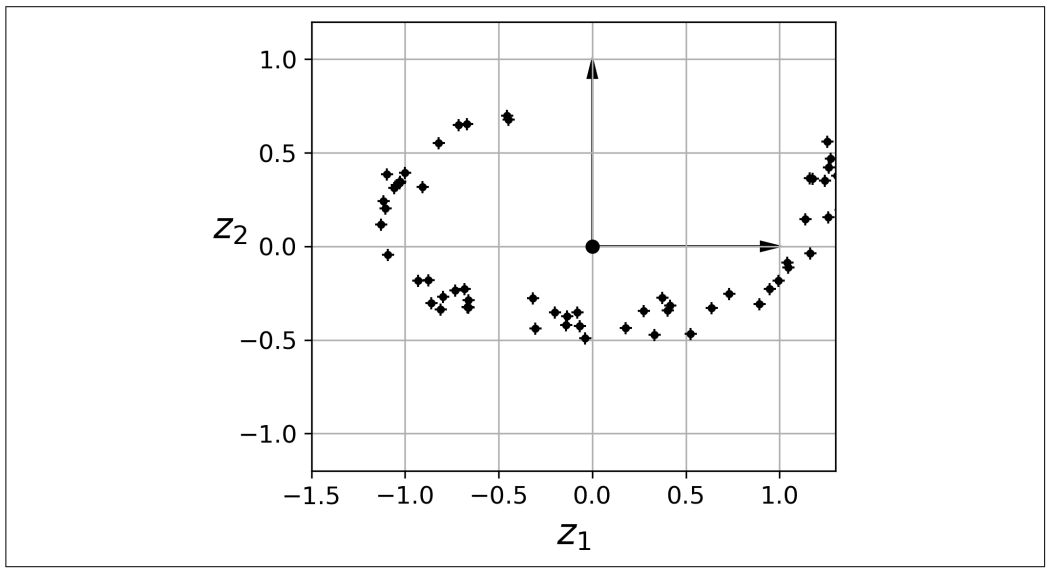


Figure 8-3. The new 2D dataset after projection

Covariance matrix

Covariance measures the strength of the linear relationship between two variables

$$\sigma_{xy} = \mathrm{E}[(x - \mu_x)(y - \mu_y)] \tag{1}$$

Covariance matrix C for multivariate random variable X

$$C_{ij} = \mathrm{E}[(x_i - \mu_i)(x_j - \mu_j)] \tag{2}$$

Principal component analysis (PCA)

Preserving the variance

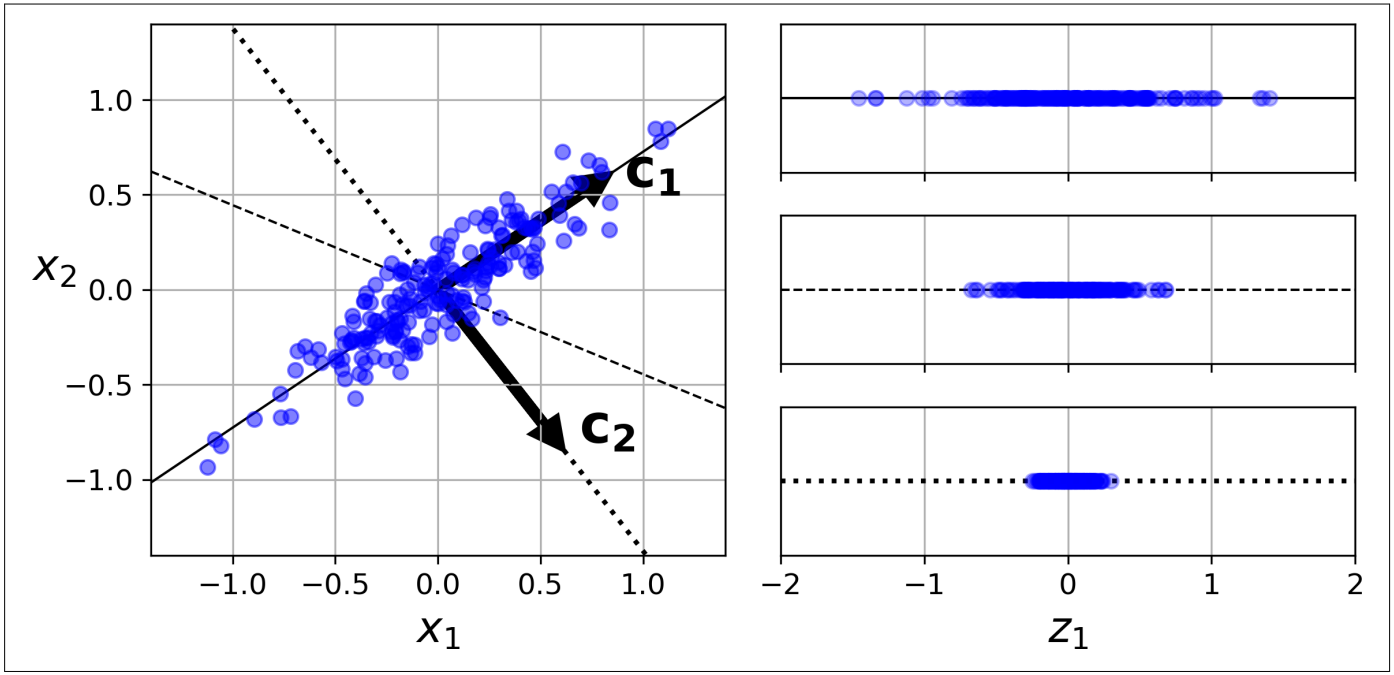


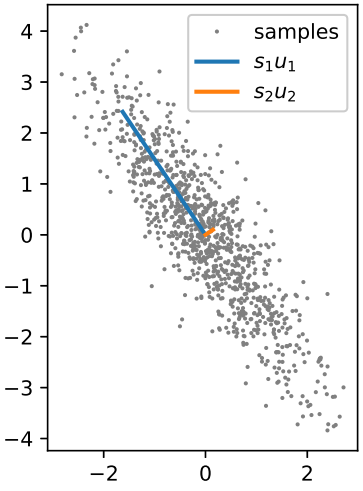
Figure 8-7. Selecting the subspace onto which to project

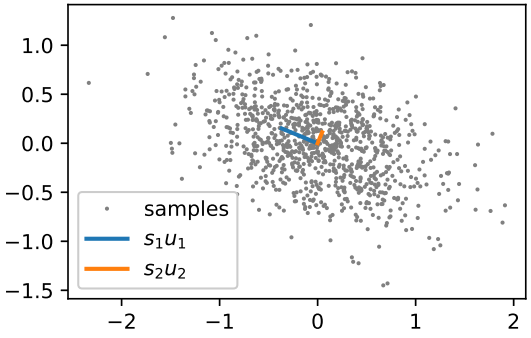
Principal component analysis (PCA)

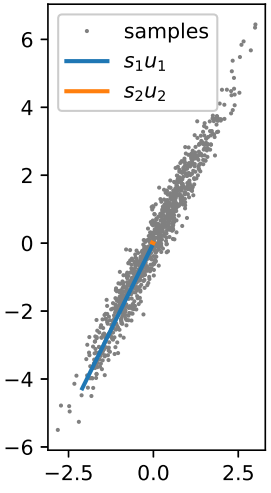
For given data $x_1, x_2, ..., x_N \in \Re^D$

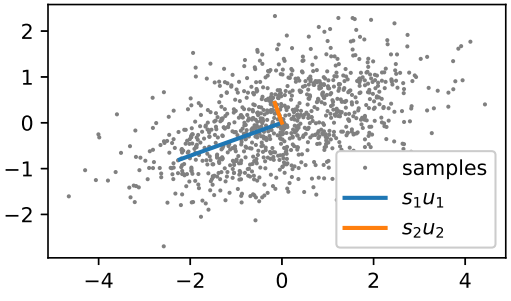
1. create a matrix $X \in \Re^{D \times N}$ with one column vector per each sample
2. covariance matrix $\Sigma = \mathrm{E}\left[(X - \mathrm{E}(X))(X - \mathrm{E}(X))^T\right] \in \Re^{D \times D}$
3. find singular vectors and singular values of Σ
4. principal components = largest singular values and vectors

Principal component analysis (PCA)









Principal component analysis (PCA)

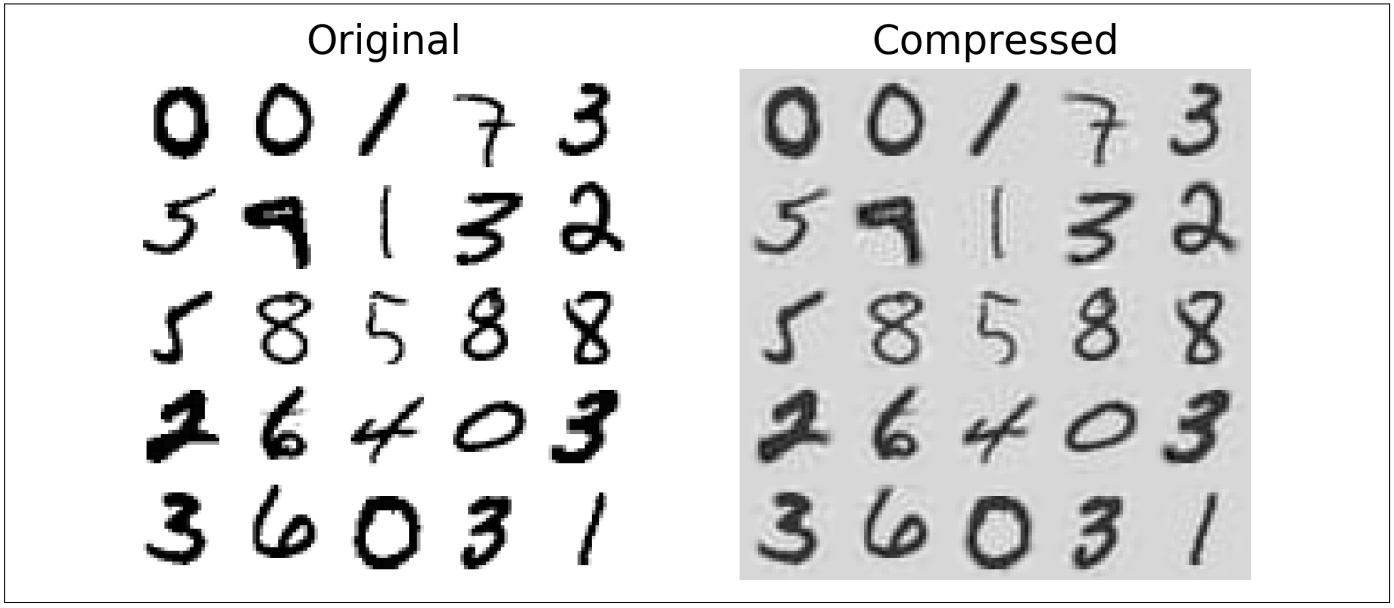


Figure 8-9. MNIST compression preserving 95% of the variance