Part I

Modeling Data with a Single Subspace

Chapter 2

Principal Component Analysis

> *"Principal component analysis is probably the oldest and best known of the techniques of multivariate analysis."*
>
> – Ian T. Jolliffe

Principal component analysis (PCA) is the problem of fitting a low-dimensional affine subspace to a set of data points in a high-dimensional space. PCA is, by now, well established in the literature, and has become one of the most useful tools for data modeling, compression, and visualization.

In this chapter, we will give a brief review of the basic principles behind PCA. When the dimension of the subspace is known, we introduce both the statistical and geometric formulations of the PCA problem and establish their equivalence. Specifically, we show that the singular value decomposition provides an optimal solution to the PCA problem and provide an interpretation of it as a rank minimization problem. We also establish the similarities and differences between PCA and a probabilistic generative subspace model called probabilistic PCA. Finally, when the dimension of the subspace is unknown, we introduce some conventional model selection methods to determine the number of principal components.

2.1 Principal Component Analysis (PCA)

Principal component analysis (PCA) refers to the problem of fitting a low-dimensional affine subspace S of dimension $d \ll D$ to a set of points $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ in a high-dimensional space $\mathbb{R}^D$. Mathematically, this problem can be formulated as either a statistical problem or a geometric one. In this section, we will discuss both formulations and show that they lead to the same solution. We will also formulate PCA as a low-rank matrix approximation problem.

2.1.1 A Statistical View of PCA

Historically, PCA was first formulated in a statistical setting to estimate the principal components of a multivariate random variable $\boldsymbol{x}$ [Pearson, 1901, Hotelling, 1933]. Specifically, given a zero-mean multivariate random variable $\boldsymbol{x} \in \mathbb{R}^D$ and any integer $d < D$, the d "principal components" of $\boldsymbol{x}$, $\boldsymbol{y} \in \mathbb{R}^d$, are defined as the d *uncorrelated* linear components of $\boldsymbol{x}$,

$$y_i = \boldsymbol{u}_i^\top \boldsymbol{x} \ \in \mathbb{R}, \quad \boldsymbol{u}_i \in \mathbb{R}^D, \quad i = 1, 2, \dots, d, \tag{2.1}$$

such that the variance of y_i is maximized subject to

$$\boldsymbol{u}_i^\top \boldsymbol{u}_i = 1 \quad \text{and} \quad \text{Var}(y_1) \geq \text{Var}(y_2) \geq \cdots \geq \text{Var}(y_d) > 0. \tag{2.2}$$

For example, to find the first principal component, y_1, we seek a vector $\boldsymbol{u}_1^* \in \mathbb{R}^D$ such that

$$\boldsymbol{u}_1^* = \arg\max_{\boldsymbol{u}_1 \in \mathbb{R}^D} \text{Var}(\boldsymbol{u}_1^\top \boldsymbol{x}) \quad \text{s.t.} \quad \boldsymbol{u}_1^\top \boldsymbol{u}_1 = 1. \tag{2.3}$$

The following theorem shows that the principal components of $\boldsymbol{x}$ can be computed from the eigenvectors of its covariance matrix $\Sigma_{\boldsymbol{x}} \doteq \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]$.

Theorem 2.1 (Principal Components of a Random Variable)**.** *Assume that* $\text{rank}(\Sigma_{\boldsymbol{x}}) \geq d$*. Then, the first d principal components of a zero-mean multivariate random variable $\boldsymbol{x}$, denoted by y_i for $i = 1, 2, \dots, d$, are given by*

$$y_i = \boldsymbol{u}_i^\top \boldsymbol{x}, \tag{2.4}$$

where $\{\boldsymbol{u}_i\}_{i=1}^d$ are d orthonormal eigenvectors of $\Sigma_{\boldsymbol{x}} \doteq \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]$ associated with its d largest eigenvalues $\{\lambda_i\}_{i=1}^d$. Moreover, $\lambda_i = \text{Var}(y_i)$ for $i = 1, 2, \dots, d$.

Proof. For the sake of simplicity, let us first assume that $\Sigma_{\boldsymbol{x}}$ does not have repeated eigenvalues. In this case, since the matrix $\Sigma_{\boldsymbol{x}}$ is real and symmetric, its eigenvalues are real and its eigenvectors form a basis of $\mathbb{R}^D$. Moreover, the eigenvectors are unique (up to sign) and the eigenvectors corresponding to different eigenvalues are orthogonal to each other (see Exercise 2.1).

Now, notice that for any $\boldsymbol{u} \in \mathbb{R}^D$, we have that

$$\text{Var}(\boldsymbol{u}^\top \boldsymbol{x}) = \mathbb{E}[(\boldsymbol{u}^\top \boldsymbol{x})^2] = \mathbb{E}[\boldsymbol{u}^\top \boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{u}] = \boldsymbol{u}^\top \Sigma_{\boldsymbol{x}} \boldsymbol{u}. \tag{2.5}$$

Therefore, the optimization problem in (2.3) is equivalent to

$$\max_{\boldsymbol{u}_1 \in \mathbb{R}^D} \boldsymbol{u}_1^\top \Sigma_{\boldsymbol{x}} \boldsymbol{u}_1 \quad \text{s.t.} \quad \boldsymbol{u}_1^\top \boldsymbol{u}_1 = 1. \tag{2.6}$$

To solve the above constrained minimization problem, we use the method of Lagrange multipliers (see Appendix A). The Lagrangian is given by

$$\mathcal{L} = \boldsymbol{u}_1^\top \Sigma_{\boldsymbol{x}} \boldsymbol{u}_1 + \lambda_1 (1 - \boldsymbol{u}_1^\top \boldsymbol{u}_1), \tag{2.7}$$

where $\lambda_1 \in \mathbb{R}$ is the Lagrange multiplier. From computing the derivatives of $\mathcal{L}$ with respect to $(\boldsymbol{u}_1, \lambda_1)$ and setting them to zero, we obtain the following necessary conditions for $(\boldsymbol{u}_1, \lambda_1)$ to be an extremum of $\mathcal{L}$:

$$\Sigma_{\boldsymbol{x}} \boldsymbol{u}_1 = \lambda_1 \boldsymbol{u}_1 \quad \text{and} \quad \boldsymbol{u}_1^\top \boldsymbol{u}_1 = 1. \tag{2.8}$$

This means that $\boldsymbol{u}_1$ is an eigenvector of $\Sigma_{\boldsymbol{x}}$ with associated eigenvalue λ_1. Since the extremum value is $\boldsymbol{u}_1^\top \Sigma_{\boldsymbol{x}} \boldsymbol{u}_1 = \lambda_1 \boldsymbol{u}_1^\top \boldsymbol{u}_1 = \lambda_1$, the optimal solution for $\boldsymbol{u}_1$ is given by the eigenvector of $\Sigma_{\boldsymbol{x}}$ associated with its largest eigenvalue $\lambda_1 = \text{Var}(y_1) > 0$.

To find the second principal component, $\boldsymbol{u}_2$, we use the fact that $\boldsymbol{u}_1^\top \boldsymbol{x}$ and $\boldsymbol{u}_2^\top \boldsymbol{x}$ need to be uncorrelated. This implies that $\boldsymbol{u}_2$ is orthogonal to $\boldsymbol{u}_1$: Indeed from

$$\mathbb{E}[(\boldsymbol{u}_1^\top \boldsymbol{x})(\boldsymbol{u}_2^\top \boldsymbol{x})] = \mathbb{E}[\boldsymbol{u}_1^\top \boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{u}_2] = \boldsymbol{u}_1^\top \Sigma_{\boldsymbol{x}} \boldsymbol{u}_2 = \lambda_1 \boldsymbol{u}_1^\top \boldsymbol{u}_2 = 0 \tag{2.9}$$

and $\lambda_1 \neq 0$, we have $\boldsymbol{u}_1^\top \boldsymbol{u}_2 = 0$. Thus, to find $\boldsymbol{u}_2$, we need to solve the following optimization problem

$$\max_{u_2 \in \mathbb{R}^D} \boldsymbol{u}_2^\top \Sigma_{\boldsymbol{x}} \boldsymbol{u}_2 \quad \text{s.t.} \quad \boldsymbol{u}_2^\top \boldsymbol{u}_2 = 1 \quad \text{and} \quad \boldsymbol{u}_1^\top \boldsymbol{u}_2 = 0. \tag{2.10}$$

As before, we define the Lagrangian

$$\mathcal{L} = \boldsymbol{u}_2^\top \Sigma_{\boldsymbol{x}} \boldsymbol{u}_2 + \lambda_2 (1 - \boldsymbol{u}_2^\top \boldsymbol{u}_2) + \gamma \boldsymbol{u}_1^\top \boldsymbol{u}_2. \tag{2.11}$$

The necessary conditions for $(\boldsymbol{u}_2, \lambda_2)$ to be an extremum are

$$\Sigma_{\boldsymbol{x}} \boldsymbol{u}_2 + \frac{\gamma}{2} \boldsymbol{u}_1 = \lambda_2 \boldsymbol{u}_2, \quad \boldsymbol{u}_2^\top \boldsymbol{u}_2 = 1 \quad \text{and} \quad \boldsymbol{u}_1^\top \boldsymbol{u}_2 = 0, \tag{2.12}$$

from which it follows that $\boldsymbol{u}_1^\top \Sigma_{\boldsymbol{x}} \boldsymbol{u}_2 + \frac{\gamma}{2} \boldsymbol{u}_1^\top \boldsymbol{u}_1 = \lambda_1 \boldsymbol{u}_1^\top \boldsymbol{u}_2 + \frac{\gamma}{2} = \lambda_2 \boldsymbol{u}_1^\top \boldsymbol{u}_2$, and so $\gamma = 2(\lambda_2 - \lambda_1) \boldsymbol{u}_1^\top \boldsymbol{u}_2 = 0$. This implies that $\Sigma_{\boldsymbol{x}} \boldsymbol{u}_2 = \lambda_2 \boldsymbol{u}_2$ and that the extremum value is $\boldsymbol{u}_2^\top \Sigma_{\boldsymbol{x}} \boldsymbol{u}_2 = \lambda_2 = \text{Var}(y_2)$. Therefore, $\boldsymbol{u}_2$ is the leading eigenvector of $\Sigma_{\boldsymbol{x}}$ restricted to the orthogonal complement of $\boldsymbol{u}_1$.[1] Since the eigenvalues of $\Sigma_{\boldsymbol{x}}$ are different, $\boldsymbol{u}_2$ is the eigenvector of $\Sigma_{\boldsymbol{x}}$ associated with its second largest eigenvalue.

To find the remaining principal components, we use that fact that for all for $i \neq j$, $y_i = u_i^\top \boldsymbol{x}$ and $y_j = \boldsymbol{u}_j^\top \boldsymbol{x}$ need to be uncorrelated, hence

$$\text{Var}(y_i y_j) = \mathbb{E}[\boldsymbol{u}_i^\top \boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{u}_j] = \boldsymbol{u}_i^\top \Sigma_{\boldsymbol{x}} \boldsymbol{u}_j = 0.$$

Using induction, assume that $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{i-1}$ are the unit-length eigenvectors of $\Sigma_{\boldsymbol{x}}$ associated with its top $i-1$ eigenvalues and let $\boldsymbol{u}_i$ be the vector defining the i-th principal component y_i. Then $\Sigma_{\boldsymbol{x}} \boldsymbol{u}_j = \lambda_j \boldsymbol{u}_j$ for $j = 1, \ldots, i-1$ and $\boldsymbol{u}_i^\top \Sigma_{\boldsymbol{x}} \boldsymbol{u}_j = \lambda_j \boldsymbol{u}_i^\top \boldsymbol{u}_j = 0$ for all $j = 1, \ldots, i-1$. Since $\lambda_j > 0$, we have that $\boldsymbol{u}_i^\top \boldsymbol{u}_j = 0$ for all $j = 1, \ldots, i-1$. To compute $\boldsymbol{u}_i$, we build the Lagrangian

$$\mathcal{L} = \boldsymbol{u}_i^\top \Sigma_{\boldsymbol{x}} \boldsymbol{u}_i + \lambda_i (1 - \boldsymbol{u}_i^\top \boldsymbol{u}_i) + \sum_{j=1}^{i-1} \gamma_j \boldsymbol{u}_i^\top \boldsymbol{u}_j. \tag{2.13}$$

[1] The reason for this is that both $\boldsymbol{u}_1$ and its orthogonal complement $\boldsymbol{u}_1^\perp$ are invariant subspaces of $\Sigma_{\boldsymbol{x}}$.

The necessary conditions for $(\boldsymbol{u}_i, \lambda_i)$ to be an extremum are

$$\Sigma_{\boldsymbol{x}}\boldsymbol{u}_i + \sum_{j=1}^{i-1} \frac{\gamma_j}{2}\boldsymbol{u}_j = \lambda_i \boldsymbol{u}_i,\ \boldsymbol{u}_i^\top \boldsymbol{u}_i = 1 \text{ and } \boldsymbol{u}_i^\top \boldsymbol{u}_j = 0, j = 1, \dots, i-1, \quad (2.14)$$

from which it follows that $\boldsymbol{u}_j^\top \Sigma_{\boldsymbol{x}} \boldsymbol{u}_i + \frac{\gamma_j}{2} = \lambda_j \boldsymbol{u}_j^\top \boldsymbol{u}_i + \frac{\gamma_j}{2} = \lambda_i \boldsymbol{u}_j^\top \boldsymbol{u}_i$, and so $\gamma_j = 2(\lambda_j - \lambda_i)\boldsymbol{u}_j^\top \boldsymbol{u}_i = 0$ for all $j = 1, \dots, i-1$. Since the associated extremum value is $\boldsymbol{u}_i^\top \Sigma_{\boldsymbol{x}} \boldsymbol{u}_i = \lambda_i = \text{Var}(y_i)$, $\boldsymbol{u}_i$ is the leading eigenvector of $\Sigma_{\boldsymbol{x}}$ restricted to the orthogonal complement of the span of $\boldsymbol{u}_1, \dots, \boldsymbol{u}_{i-1}$. Since the eigenvalues of $\Sigma_{\boldsymbol{x}}$ are different, $\boldsymbol{u}_i$ is the eigenvector of $\Sigma_{\boldsymbol{x}}$ associated with the i-th largest eigenvalue. Therefore, when the eigenvalues of $\Sigma_{\boldsymbol{x}}$ are different, each eigenvector $\boldsymbol{u}_i$ is unique (up to sign), hence so are the principal components of $\boldsymbol{x}$.

Let us now consider the case where $\Sigma_{\boldsymbol{x}}$ has repeated eigenvalues. In this case, $\Sigma_{\boldsymbol{x}}$ still admits a basis of orthonormal eigenvectors. Specifically, the eigenvectors of $\Sigma_{\boldsymbol{x}}$ associated to different eigenvalues are still orthogonal, while the eigenvectors associated with a repeated eigenvalue form an eigensubspace and any orthonormal basis for this eigensubspace gives a valid set of eigenvectors (see Exercise 2.1). As a consequence, the principal directions $\{\boldsymbol{u}_i\}_{i=1}^d$ are not uniquely defined. For example, if λ_1 is repeated, any eigenvector associated with λ_1 can be chosen as $\boldsymbol{u}_1$ and any other eigenvector associated with λ_1 and orthogonal to $\boldsymbol{u}_1$ can be chosen as $\boldsymbol{u}_2$. Nonetheless, the principal components can still be obtained from a(ny) set of the top d eigenvectors of $\Sigma_{\boldsymbol{x}}$, as claimed. □

The solution to PCA provided by Theorem 2.1 suggests that we may find the d principal components of $\boldsymbol{x}$ simultaneously, rather than one by one. Specifically, we can define a random vector $\boldsymbol{y} = [y_1, y_2, \dots, y_d]^\top \in \mathbb{R}^d$ and a matrix $U_d = [\boldsymbol{u}_1, \boldsymbol{u}_2, \dots, \boldsymbol{u}_d] \in \mathbb{R}^{D \times d}$. Since $\boldsymbol{y} = U_d^\top \boldsymbol{x}$, we have that

$$\Sigma_{\boldsymbol{y}} \doteq \mathbb{E}[\boldsymbol{y}\boldsymbol{y}^\top] = U_d^\top \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] U_d = U_d^\top \Sigma_{\boldsymbol{x}} U_d. \quad (2.15)$$

Since were are looking for uncorrelated random variables, the matrix $\Sigma_{\boldsymbol{y}}$ must be diagonal and the matrix U_d must be orthonormal, i.e., $U_d^\top U_d = I_d$.

Recall that any diagonalizable matrix A can be transformed into a diagonal matrix $\Lambda = V^{-1}AV$, where the columns of V are the eigenvectors of A and the diagonal entries of Λ are the corresponding eigenvalues. Recall also that if A is real, symmetric and positive semi-definite, its eigenvalues are real and nonnegative, i.e., $\lambda_i \geq 0$, and its eigenvectors can be chosen to be orthonormal, so that $V^{-1} = V^\top$ (see Exercise 2.1). Since the matrix $\Sigma_{\boldsymbol{x}}$ is real, symmetric and positive semi-definite, one solution to the equation $\Sigma_{\boldsymbol{y}} = U_d^\top \Sigma_{\boldsymbol{x}} U_d$ is obtained by choosing the columns of U_d as d eigenvectors of $\Sigma_{\boldsymbol{x}}$ and the diagonal entries of $\Sigma_{\boldsymbol{y}}$ as the corresponding d eigenvalues. Moreover, since our goal is to maximize the variance of each y_i and $\lambda_i = \text{Var}(y_i)$, we conclude that the columns of U_d are the top d eigenvectors of $\Sigma_{\boldsymbol{x}}$ and the entries of $\Sigma_{\boldsymbol{y}}$ are the corresponding top d eigenvalues.

Principal Components of a Non-zero Mean Random Variable

When $\boldsymbol{x}$ is not zero mean, the d principal components of $\boldsymbol{x}$ are defined as the d uncorrelated affine components of $\boldsymbol{x}$

$$y_i = \boldsymbol{u}_i^\top \boldsymbol{x} + a_i \ \in \mathbb{R}, \quad \boldsymbol{u}_i \in \mathbb{R}^D, \quad i = 1, 2, \ldots, d, \tag{2.16}$$

such that the variance of y_i is maximized subject to

$$\boldsymbol{u}_i^\top \boldsymbol{u}_i = 1 \quad \text{and} \quad \text{Var}(y_1) \geq \text{Var}(y_2) \geq \cdots \geq \text{Var}(y_d) > 0. \tag{2.17}$$

As shown in Exercise 2.3, the principal directions $\{\boldsymbol{u}_i\}_{i=1}^d$ are the d eigenvectors of $\Sigma_{\boldsymbol{x}} \doteq \mathbb{E}[(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^\top]$, where $\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{x})$, associated with its d largest eigenvalues $\{\lambda_i\}_{i=1}^d$. Moreover, $\lambda_i = \text{Var}(y_i)$ and $a_i = -\boldsymbol{u}_i^\top \boldsymbol{\mu}$ for $i = 1, 2, \ldots, d$.

Sample Principal Components of a Zero Mean Random Variable

In practice, we may not know the population covariance matrix, $\Sigma_{\boldsymbol{x}}$. Instead, we may be given N i.i.d. samples of the zero-mean random variable $\boldsymbol{x}$, $\{\boldsymbol{x}_j\}_{j=1}^N$, which we collect into a data matrix $X = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N]$. It is well known from statistics (see Exercise B.1) that the maximum likelihood estimate of $\Sigma_{\boldsymbol{x}}$ is given by

$$\widehat{\Sigma}_N \doteq \frac{1}{N} \sum_{j=1}^N \boldsymbol{x}_j \boldsymbol{x}_j^\top = \frac{1}{N} X X^\top. \tag{2.18}$$

We define the d "sample principal components" of $\boldsymbol{x}$ as

$$\widehat{y}_i = \widehat{\boldsymbol{u}}_i^\top \boldsymbol{x}, \quad i = 1, 2, \ldots, d, \tag{2.19}$$

where $\{\widehat{\boldsymbol{u}}_i\}_{i=1}^d$ are the top d eigenvectors of $\widehat{\Sigma}_N$, or equivalently those of $XX^\top$.

Notice that when the dimension of the data, D, is very high, we can avoid computing the eigenvectors of a large matrix $XX^\top$ by exploiting the fact that the top eigenvectors of $XX^\top$ are the same as the top singular vectors of X. Therefore, the sample principal components of $\boldsymbol{x}$ may be computed from the singular value decomposition (SVD) of $X = U \Sigma V^\top$ as $\boldsymbol{y} = U_d^\top \boldsymbol{x}$, where the columns of U_d are the first d columns of U.

Remark 2.2 (Relationship between principal components and sample principal components)**.** *Even though the principal components of $\boldsymbol{x}$ and the sample principal components of $\boldsymbol{x}$ are different notions, under certain assumptions on the distribution of $\boldsymbol{x}$ they can be related to each other. Specifically, one can show that, if $\boldsymbol{x}$ is Gaussian, then every eigenvector $\widehat{\boldsymbol{u}}$ of $\widehat{\Sigma}_N$ is an asymptotically consistent unbiased estimate (see Appendix B) for the corresponding eigenvector $\boldsymbol{u}$ of $\Sigma_{\boldsymbol{x}}$. Interested readers may find a more detailed proof in [Jolliffe, 1986b].*

2.1.2 A Geometric View of PCA

An alternative geometric view of PCA, which is very much related to the SVD [Beltrami, 1873,Jordan, 1874], assumes that we are given a set of points $\{\boldsymbol{x}_j\}_{j=1}^N$ in $\mathbb{R}^D$ and seeks to find an (affine) subspace $S \subset \mathbb{R}^D$ of dimension d that best fits these points. Each point $\boldsymbol{x}_j \in S$ can be represented as

$$\boldsymbol{x}_j = \boldsymbol{\mu} + U_d \boldsymbol{y}_j, \quad j = 1, 2, \dots, N, \tag{2.20}$$

where $\boldsymbol{\mu} \in S$ is a(ny) point in the subspace, U_d is a $D \times d$ matrix whose columns form a basis for the subspace, and $\boldsymbol{y}_j \in \mathbb{R}^d$ is simply the vector of new coordinates of $\boldsymbol{x}_j$ in the subspace.

Notice that there is some redundancy in the above representation due to the arbitrariness in the choice of $\boldsymbol{\mu}$ and U_d. More precisely, for any $\boldsymbol{y}_0 \in \mathbb{R}^d$, we can re-represent $\boldsymbol{x}_j$ as $\boldsymbol{x}_j = (\boldsymbol{\mu} + U_d\boldsymbol{y}_0) + U_d(\boldsymbol{y}_j - \boldsymbol{y}_0)$. We call this ambiguity the *translational ambiguity*. Also, for any $A \in \mathbb{R}^{d\times d}$ we can re-represent $\boldsymbol{x}_j$ as $\boldsymbol{x}_j = \boldsymbol{\mu} + (U_d A)(A^{-1}\boldsymbol{y}_j)$. We call this ambiguity the *change of basis ambiguity*. Therefore, we need some additional constraints in order to end up with a unique solution to the problem of finding an affine subspace for the data.

A common constraint used to resolve the translational ambiguity is to impose that the average of the $\boldsymbol{y}_j$ be zero,[2] i.e.,

$$\frac{1}{N}\sum_{j=1}^{N} \boldsymbol{y}_j = \boldsymbol{0}, \tag{2.21}$$

where $\boldsymbol{0} \in \mathbb{R}^d$ is the vector of all zeros, while a common constraint used to resolve the change of basis ambiguity is to impose that the columns of U_d be orthonormal, i.e., $U_d^\top U_d = I$. This last constraint eliminates the change of basis ambiguity only up to a rotation, because we can still re-represent $\boldsymbol{x}_j$ as $\boldsymbol{x}_j = \boldsymbol{\mu} + (U_d R)(R^\top \boldsymbol{y}_j)$ for some rotation R in $\mathbb{R}^d$. However, this *rotational ambiguity* can easily be dealt with during optimization, as we shall see.

In general the given points are imperfect and have noise. For example, if the points N are contaminated by additive noise $\{\varepsilon_j\}_{j=1}^N$, respectively, we have that

$$\boldsymbol{x}_j = \boldsymbol{\mu} + U_d \boldsymbol{y}_j + \varepsilon_j, \quad j = 1, 2, \dots, N. \tag{2.22}$$

In this case, we define the "optimal" affine subspace to be the one that minimizes the sum of squared errors, i.e.,

$$\min_{\boldsymbol{\mu}, U_d, \{\boldsymbol{y}_j\}} \sum_{j=1}^{N} \left\| \boldsymbol{x}_j - \boldsymbol{\mu} - U_d \boldsymbol{y}_j \right\|^2, \ \text{ s.t. } \ U_d^\top U_d = I_d \ \text{ and } \ \sum_{j=1}^{N} \boldsymbol{y}_j = \boldsymbol{0}. \tag{2.23}$$

[2] In the statistical setting, $\boldsymbol{x}_j$ and $\boldsymbol{y}_j$ will be samples of two random variables $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. Then this constraint is equivalent to setting their means to be zero.

In order to solve this optimization problem, we define the Lagrangian

$$\mathcal{L} = \sum_{j=1}^{N} \left\| \boldsymbol{x}_j - \boldsymbol{\mu} - U_d \boldsymbol{y}_j \right\|^2 + \gamma^\top \sum_{j=1}^{N} \boldsymbol{y}_j + \operatorname{trace}\big((I_d - U_d^\top U_d)\Lambda\big), \tag{2.24}$$

where $\gamma \in \mathbb{R}^d$ and $\Lambda = \Lambda^\top \in \mathbb{R}^{d \times d}$ are, respectively, a vector and a matrix of Lagrange multipliers. The necessary condition for $\boldsymbol{\mu}$ to be an extremum is

$$-2 \sum_{j=1}^{N} (\boldsymbol{x}_j - \boldsymbol{\mu} - U_d \boldsymbol{y}_j) = \mathbf{0} \implies \widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{\mu}}_N \doteq \frac{1}{N} \sum_{j=1}^{N} \boldsymbol{x}_j. \tag{2.25}$$

The necessary condition for $\boldsymbol{y}_j$ to be an extremum is

$$-2U_d^\top (\boldsymbol{x}_j - \boldsymbol{\mu} - U_d \boldsymbol{y}_j) + \gamma = \mathbf{0}. \tag{2.26}$$

Summing over j yields $\gamma = \mathbf{0}$, from which we obtain

$$\widehat{\boldsymbol{y}}_j = U_d^\top (\boldsymbol{x}_j - \widehat{\boldsymbol{\mu}}_N). \tag{2.27}$$

The vector $\widehat{\boldsymbol{y}}_j \in \mathbb{R}^d$ is simply the coordinates of the projection of $\boldsymbol{x}_j \in \mathbb{R}^D$ onto the subspace S. We may call such $\widehat{\boldsymbol{y}}$ the "geometric principal components" of $\boldsymbol{x}$.

Before optimizing over U_d, we can replace the optimal values for $\boldsymbol{\mu}$ and $\boldsymbol{y}_j$ into the objective function. This leads to the following optimization problem

$$\min_{U_d} \sum_{j=1}^{N} \left\| (\boldsymbol{x}_j - \widehat{\boldsymbol{\mu}}_N) - U_d U_d^\top (\boldsymbol{x}_j - \widehat{\boldsymbol{\mu}}_N) \right\|^2 \quad \text{s.t.} \quad U_d^\top U_d = I_d. \tag{2.28}$$

Note that this is a restatement of the original problem with the mean $\widehat{\boldsymbol{\mu}}_N$ subtracted from each of the sample points. Therefore, from now on, we will consider only the case in which the data points have zero mean. If not, simply subtract the mean from each point before computing U_d.

The following theorem gives a constructive solution for finding an optimal $\widehat{U}_d$.

Theorem 2.3 (PCA via SVD). *Let $X = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N] \in \mathbb{R}^{D \times N}$ be the matrix formed by stacking the (zero-mean) data points as its column vectors. Let $X = U\Sigma V^\top$ be the SVD of the matrix X. Then for any given $d < D$, an optimal solution $\widehat{U}_d$ for U_d is given by the first d columns of U, and an optimal solution $\widehat{\boldsymbol{y}}_j$ for $\boldsymbol{y}_j$ is given by the i-th column of the top $d \times N$ submatrix $\widehat{\Sigma}_d \widehat{V}_d^\top$ of $\Sigma V^\top$.*

Proof. Recalling that $\boldsymbol{x}^\top A \boldsymbol{x} = \operatorname{trace}(A\boldsymbol{x}\boldsymbol{x}^\top)$, we can rewrite the least-squares error

$$\sum_{j=1}^{N} \left\| \boldsymbol{x}_j - U_d U_d^\top \boldsymbol{x}_j \right\|^2 = \sum_{j=1}^{N} \boldsymbol{x}_j^\top (I_D - U_d U_d^\top) \boldsymbol{x}_j \tag{2.29}$$

as $\operatorname{trace}((I_D - U_d U_d^\top) X X^\top)$. The first term $\operatorname{trace} X X^\top$ does not depend on U_d. Therefore, we can transform the minimization of (2.29) to

$$\max_{U_d} \quad \operatorname{trace}(U_d U_d^\top X X^\top) \quad \text{s.t.} \quad U_d^\top U_d = I_d. \tag{2.30}$$

Since $\operatorname{trace} AB = \operatorname{trace} BA$, the Lagrangian for this problem can be written as

$$\mathcal{L} = \operatorname{trace}(U_d^\top X X^\top U_d) + \operatorname{trace}((I_d - U_d^\top U_d)\Lambda), \tag{2.31}$$

where $\Lambda = \Lambda^\top \in \mathbb{R}^{d \times d}$. The conditions for an extremum are given by

$$X X^\top U_d = U_d \Lambda. \tag{2.32}$$

Therefore, $\Lambda = U_d^\top X X^\top U_d$ and the objective function reduces to $\operatorname{trace}(\Lambda)$. Now, recall that U_d is defined only up to a rotation, i.e., $U_d' = U_d R$ is also a valid solution, hence so is $\Lambda' = R \Lambda R^\top$. Since Λ is symmetric, it has an orthogonal matrix of eigenvectors. Thus, if we choose R to be the matrix of eigenvectors of Λ, then Λ' is a diagonal matrix. As a consequence, we can choose Λ to be diagonal without loss of generality. It follows from (2.32) that the columns of U_d must be d eigenvectors of $X X^\top$ with the corresponding eigenvalues in the diagonal entries of Λ. Since the goal is to maximize $\operatorname{trace}(\Lambda)$, an optimal solution is given by the top d eigenvectors of $X X^\top$, i.e., the top d singular vectors of $X = U \Sigma V^\top$, which are the first d columns of U. It then follows from (2.27) that $Y = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N] = U_d^\top X = U_d^\top U \Sigma V^\top = \Sigma_d V_d^\top$. Finally, since $\Lambda = U_d^\top U \Sigma^2 U^\top U_d = \Sigma_d^2$, the optimal least-squares error is given by $\operatorname{trace}(\Sigma^2) - \operatorname{trace}(\Sigma_d^2) = \sum_{i=d+1}^D \sigma_i^2$, where σ_i is the i-th singular value of X. □

According to the theorem, the SVD gives an optimal solution to the PCA problem. The resulting matrix $\widehat{U}_d$ (together with the mean $\widehat{\boldsymbol{\mu}}$ if the data is not zero-mean) provides a geometric description of the dominant subspace structure for all the points;[3] and the columns of the matrix $\widehat{\Sigma}_d \widehat{V}_d^\top = [\widehat{\boldsymbol{y}}_1, \widehat{\boldsymbol{y}}_2, \ldots, \widehat{\boldsymbol{y}}_N] \in \mathbb{R}^{d \times N}$, i.e., the principal components, give a more compact representation for the points $X = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N] \in \mathbb{R}^{D \times N}$, as d is typically much smaller than D.

Theorem 2.4 (Equivalence of Geometric and Sample Principal Components)**.** *Let $X = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N] \in \mathbb{R}^{D \times N}$ be the mean-subtracted data matrix. The vectors $\widehat{\boldsymbol{u}}_1, \widehat{\boldsymbol{u}}_2, \ldots, \widehat{\boldsymbol{u}}_d \in \mathbb{R}^D$ associated with the d sample principal components of X are exactly the columns of the matrix $\widehat{U}_d \in \mathbb{R}^{D \times d}$ that minimizes the least-squares error (2.29).*

Proof. The proof is simple. Notice that if X has the singular value decomposition $X = U \Sigma V^\top$, then $X X^\top = U \Sigma^2 U^\top$ is the eigenvalue decomposition of $X X^\top$. If Σ is ordered, then the first d columns of U are exactly the leading d eigenvectors of $X X^\top$, which give the d sample principal components. □

The above theorem shows that both the geometric and statistical formulations of PCA lead to exactly the same solution/estimate of the sample principal components. This equivalence is part of the reason why PCA has become the tool of choice for dimensionality reduction as the optimality of the solution can be interpreted either statistically or geometrically in different application contexts.

[3] From a statistical standpoint, the column vectors of U_d give the directions in which the data X has the largest variance, hence the name "principal components."

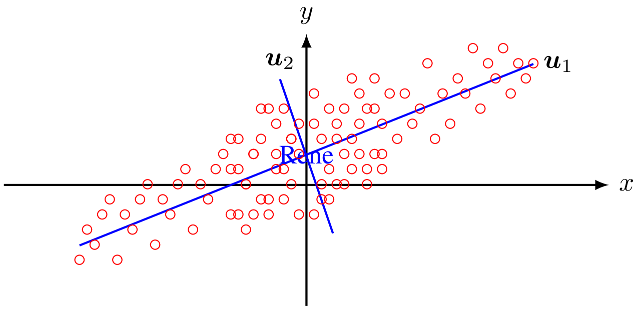

Figure 2.1. Example showing a two-dimensional dataset and its two principal components.

Figure 2.1 gives and example of a two-dimensional dataset and its two principal components.

2.1.3 A Rank Minimization View of PCA

Notice that the geometric PCA problem in (2.23) can be rewritten as

$$\min_{\boldsymbol{\mu},U_d,Y} \left\|X - \boldsymbol{\mu}\mathbf{1}^\top - U_d Y\right\|_F^2, \quad \text{s.t. } U_d^\top U_d = I_d \text{ and } Y\mathbf{1} = \mathbf{0}, \tag{2.33}$$

where $X = \left[\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\right]$, $Y = \left[\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\right]$, $\mathbf{1} \in \mathbb{R}^N$ is the vector of all ones, and $\|X\|_F^2 = \sum_{ij} X_{ij}^2$ is the Frobenius norm of X. Therefore, another interpretation of PCA is to see it as the problem of finding a vector $\boldsymbol{\mu}$ and rank-d matrix A that best approximate the data matrix X. This problem can be formulated as

$$\min_{\boldsymbol{\mu},A} \|X - \boldsymbol{\mu}\mathbf{1}^\top - A\|_F^2 \quad \text{s.t.} \quad \operatorname{rank}(A) = d \text{ and } A\mathbf{1} = \mathbf{0}. \tag{2.34}$$

Notice that this formulation is identical to that in (2.23), except that we have now replaced the subspace basis U_d and the matrix of principal components Y by their product $A = U_d Y$. The constraint $A\mathbf{1} = \mathbf{0}$ comes from the requirement that the principal components be centered, i.e., $\sum \boldsymbol{y}_j = 0$, hence $Y\mathbf{1} = \mathbf{0}$.

Since the problem in (2.34) is the same as that in (2.23), we already know that the optimal solution for $\boldsymbol{\mu}$ is $\frac{1}{N}\sum_j \boldsymbol{x}_j = \frac{1}{N}X\mathbf{1}$. Therefore, after centering the data matrix by subtracting $\boldsymbol{\mu}$ from each column, the optimization problem in (2.34) can be reduced to

$$\min_{A} \|X - A\|_F^2 \quad \text{s.t.} \quad \operatorname{rank}(A) = d. \tag{2.35}$$

Notice that we have dropped the constraint $A\mathbf{1} = \mathbf{0}$. This is because this constraint is not needed when the data matrix is centered, i.e., when $X\mathbf{1} = \mathbf{0}$. To see this, let $A^* = \arg\min_{A: A\mathbf{1}=\mathbf{0}, \operatorname{rank}(A)=d} \|X - A\|_F^2$ be the optimal solution subject to the constraint $A\mathbf{1} = \mathbf{0}$, and let $\hat{A} = A^* - \boldsymbol{a}\mathbf{1}^\top$, where $\boldsymbol{a} = \frac{1}{N}A^*\mathbf{1}$, be another solution. By the sake of contradiction, assume that $\boldsymbol{a} \neq \boldsymbol{0}$. Then,

$$\begin{aligned}\|X - \hat{A}\|_F^2 &= \|X - A^* + \boldsymbol{a}\boldsymbol{1}^\top\|_F^2 && (2.36)\\ &= \|X - A^*\|_F^2 + 2\langle X - A^*, \boldsymbol{a}\boldsymbol{1}^\top\rangle + \|\boldsymbol{a}\boldsymbol{1}^\top\|_F^2 && (2.37)\\ &= \|X - A^*\|_F^2 + 2\boldsymbol{a}^\top(X - A^*)\boldsymbol{1} + N\|\boldsymbol{a}\|_2^2 && (2.38)\\ &= \|X - A^*\|_F^2 - N\|\boldsymbol{a}\|_2^2 < \|X - A^*\|_F^2, && (2.39)\end{aligned}$$

which contradicts the optimality of A^*.

To solve the problem in (2.35), let $X = U_X \Sigma_X V_X^\top$ and $A = U_A \Sigma_A V_A^\top$ be, respectively, the SVD of X and A. Then, letting $U = U_X^\top U_A$ and $V = V_X^\top V_A$, we have

$$\begin{aligned}\|X - A\|_F^2 = \|U_X\Sigma_X V_X^\top - U_A\Sigma_A V_A^\top\|_F^2 &= \|\Sigma_X - U\Sigma_A V^\top\|_F^2 && (2.40)\\ &= \|\Sigma_X\|_F^2 - 2\langle \Sigma_X, U\Sigma_A V^\top\rangle + \|\Sigma_A\|_F^2. && (2.41)\end{aligned}$$

Therefore, minimizing $\|X - A\|_F^2$ w.r.t. A, is equivalent to minimizing the above expression with respect to U, V and Σ_A. We will solve this problem in two steps.

In the first step, we will minimize w.r.t. U and V only. Notice that this is equivalent to.

$$\max_{U,V} \langle \Sigma_X, U\Sigma_A V^\top\rangle \tag{2.42}$$

The solution to this problem can be found from the Von Neumann's inequality, which is stated next.

Lemma 2.5 (Von Neumann's Inequality). *For any $m \times n$ real valued matrices F and G, let $\sigma_1(F) \geq \sigma_2(F) \geq \cdots \geq 0$ and $\sigma_1(G) \geq \sigma_2(G) \geq \cdots \geq 0$ be the descending singular values of F and G respectively. Then*

$$\langle F, G\rangle = \operatorname{trace}(F^\top G) \leq \sum_{i=1}^n \sigma_i(F)\sigma_i(G). \tag{2.43}$$

The case of equality occurs if and only if it is possible to find unitary matrices U_F and V_F that simultaneously singular value-decompose *F and G in the sense that*

$$F = U_F \Sigma_F V_F^\top \quad \text{and} \quad G = U_F \Sigma_G V_F^\top, \tag{2.44}$$

where Σ_F and Σ_G denote the $m \times n$ diagonal matrices with the singular values of F and G, respectively, down in the diagonal.

Proof. See [Mirsky, 1975]. □

Applying this lemma to $F = \Sigma_X$ and $G = U\Sigma_A V^\top$, we obtain

$$\langle \Sigma_X, U\Sigma_A V^\top\rangle \leq \sum_{i=1}^d \sigma_i(X)\sigma_i(A), \tag{2.45}$$

because $\sigma_i(A) = 0$ for $i > d$. Notice also that equality is achieved for example for $U = I$ and $V = I$, hence $U_A = U_X$ and $V_A = V_X$.

In the second step, we will substitute the above solutions for U and V into the objective function $\|X - A\|_F^2$ and optimize over Σ_A. We obtain the following optimization problem

$$\min_{\Sigma_A} \sum_{i=1}^{d} \sigma_i(A)^2 - 2\sum_{i=1}^{d} \sigma_i(X)\sigma_i(A). \tag{2.46}$$

Taking the derivatives with respect to $\sigma_i(A)$ and setting them to zero gives us $\sigma_i(A) = \sigma_i(X)$ for $i = 1, \ldots, d$. We thus have the following result.

Theorem 2.6. *Let $X = U\Sigma V^\top$ be the singular value decomposition of the mean subtracted data matrix. An optimal solution for the optimization problem*

$$\min_{A} \|X - A\|_F^2 \quad \textit{s.t.} \quad \operatorname{rank}(A) = d \tag{2.47}$$

is given by $A = U_d \Sigma_d V_d^\top$, where U_d, Σ_d and V_d correspond to the top d singular vectors and singular values in U, Σ and V, respectively.

Notice that this theorem is essentially equivalent to Theorem 2.3 and that the above derivation based on Von Neummann's inequality provides an alternative proof for the theorem.

In summary, we can view the PCA problem either as a statistical problem, or as a geometrical problem, or as a rank minimization problem, and all three interpretations lead to the same solution.

2.2 Probabilistic PCA (PPCA)

The PCA model described so far allows us to find a low-dimensional representation $\{\boldsymbol{y}_j \in \mathbb{R}^d\}$ of a set of sample points $\{\boldsymbol{x}_j \in \mathbb{R}^D\}$, with $d \ll D$. However, the PCA model is not a proper generative model, because the low-dimensional representation $\{\boldsymbol{y}_j\}$ and the error $\{\varepsilon_j\}$ are not treated as random variables. As a consequence, the PCA model cannot be used to generate new samples $\boldsymbol{x}$.

To address this issue, assume that the low-dimensional representation $\boldsymbol{y}$ and the error ε are independent random variables with pdfs $p(\boldsymbol{y})$ and $p(\varepsilon)$, respectively. This allows us to generate a new sample of $\boldsymbol{x}$ from samples of $\boldsymbol{y}$ and ε as

$$\boldsymbol{x} = \boldsymbol{\mu} + U_d \boldsymbol{y} + \varepsilon. \tag{2.48}$$

Assume that the mean and covariance of $\boldsymbol{y}$ are denoted as $\boldsymbol{\mu}_{\boldsymbol{y}}$ and $\Sigma_{\boldsymbol{y}}$, respectively. Assume also that ε is zero mean with covariance Σ_ε. The mean and covariance of the observations are then given by

$$\boldsymbol{\mu}_{\boldsymbol{x}} = \boldsymbol{\mu} + U_d \boldsymbol{\mu}_{\boldsymbol{y}} \quad \text{and} \quad \Sigma_{\boldsymbol{x}} = U_d \Sigma_{\boldsymbol{y}} U_d^\top + \Sigma_\varepsilon. \tag{2.49}$$

Notice that, different from the PCA problem studied in the previous section, here we no longer need to assume that U_d is a unitary matrix. This is because, once we enforce a specific type of probability distribution for $\boldsymbol{y}$, we should be able to estimate via the Maximum Likelihood (ML) principle (see Appendix B.1.4) an optimal model from the observations $\boldsymbol{x}$ without any additional constraints on the matrix U_d. The remainder of the section discusses different methods for estimating the parameters of this model, including $\boldsymbol{\mu}$, U_d, $\boldsymbol{\mu}_{\boldsymbol{y}}$, $\Sigma_{\boldsymbol{y}}$ and Σ_ε, from the mean and covariance of the population, $\boldsymbol{\mu}_{\boldsymbol{x}}$ and $\Sigma_{\boldsymbol{x}}$, or from i.i.d. samples $\{\boldsymbol{x}_j\}_{j=1}^N$.

2.2.1 PPCA from Population Mean and Covariance

Observe that, in general, we cannot uniquely recover the model parameters from $\boldsymbol{\mu}_{\boldsymbol{x}}$ and $\Sigma_{\boldsymbol{x}}$. For instance, notice that $\boldsymbol{\mu}$ and $\boldsymbol{\mu}_{\boldsymbol{y}}$ cannot be uniquely recovered from $\boldsymbol{\mu}_{\boldsymbol{x}}$. Similarly to what we did in the case of PCA, this issue can easily be resolved by assuming that $\boldsymbol{\mu}_{\boldsymbol{y}} = \mathbf{0}$. This leads to the following estimate of $\boldsymbol{\mu}$

$$\widehat{\boldsymbol{\mu}} = \boldsymbol{\mu}_{\boldsymbol{x}}, \tag{2.50}$$

which is the same estimate as that of PCA (see Exercise 2.3).

Another ambiguity that cannot be resolved in a straightforward manner is that $\Sigma_{\boldsymbol{y}}$ and Σ_ε cannot be uniquely recovered from $\Sigma_{\boldsymbol{x}}$. For instance, $\Sigma_{\boldsymbol{y}} = 0$ and $\Sigma_\varepsilon = \Sigma_{\boldsymbol{x}}$ is a valid solution. However, this solution is not meaningful, because it assigns all the information in $\Sigma_{\boldsymbol{x}}$ to the error, rather than to the low-dimensional representation. Intuitively we would like $\Sigma_{\boldsymbol{y}}$ to capture as much information about $\Sigma_{\boldsymbol{x}}$ as possible. Thus it makes sense for $\Sigma_{\boldsymbol{y}}$ to be full rank and for Σ_ε to be as close to zero as possible. Probabilistic PCA (PPCA) resolves this ambiguity by assuming that

1. The low-dimensional representation has unit covariance, i.e., $\Sigma_{\boldsymbol{y}} = I_d \in \mathbb{R}^{d\times d}$.
2. The noise covariance matrix $\Sigma_\varepsilon \in \mathbb{R}^{D\times D}$ is isotropic, i.e., $\Sigma_\varepsilon = \sigma^2 I_D$.

These assumptions lead to the following relationship

$$\Sigma_{\boldsymbol{x}} = U_d U_d^\top + \sigma^2 I_D, \tag{2.51}$$

from which it follows that the off-diagonal entries of $\Sigma_{\boldsymbol{x}}$ are equal to the off-diagonal entries of $U_d U_d^\top$. As a consequence, even though both PPCA and PCA try to capture as much information from $\Sigma_{\boldsymbol{x}}$ into $\Sigma_{\boldsymbol{y}}$, the information they attempt to capture is not the same. On the one hand, PPCA tries to find a matrix U_d such that the covariances are preserved, i.e., the off-diagonal entries of $\Sigma_{\boldsymbol{x}}$. On the other hand, PCA tries to preserve the variances, i.e., the diagonal entries of $\Sigma_{\boldsymbol{x}}$.

As it turns out, the parameters U_d and σ of the PPCA model can be computed in closed form from the SVD of the population covariance $\Sigma_{\boldsymbol{x}}$, as stated by the following theorem. Again, we emphasize that in the PPCA model, the matrix U_d can be an arbitrary matrix and it does not need to be unitary.

Theorem 2.7 (PPCA from Population Mean and Covariance)**.** *The parameters $\boldsymbol{\mu}$, U_d and σ of the PPCA model can be estimated from the population mean and covariance, $\boldsymbol{\mu}_{\boldsymbol{x}}$ and $\Sigma_{\boldsymbol{x}}$, respectively, as*

$$\widehat{\boldsymbol{\mu}} = \boldsymbol{\mu}_{\boldsymbol{x}}, \ \ \widehat{U}_d = U_1(\Lambda_1 - \widehat{\sigma}^2 I)^{1/2} R, \ \ \widehat{\sigma}^2 = \lambda_{d+1} = \lambda_{d+2} = \cdots = \lambda_D, \tag{2.52}$$

where U_1 is the matrix with the top d eigenvectors of $\Sigma_{\boldsymbol{x}}$, Λ_1 is the diagonal matrix in $\mathbb{R}^{d \times d}$ of the corresponding top d eigenvalues, $R \in \mathbb{R}^{d \times d}$ is an arbitrary orthogonal matrix and λ_i is the i-th eigenvalue of $\Sigma_{\boldsymbol{x}}$.

Proof. We have already shown in (2.50) that $\widehat{\boldsymbol{\mu}} = \boldsymbol{\mu}_{\boldsymbol{x}}$. To find σ, notice that the eigenvalues of $\Sigma_{\boldsymbol{x}}$ must be equal to the eigenvalues of $U_d U_d^\top$ plus σ^2. Since $U_d U_d^\top$ has rank d and is positive semidefinite, $D - d$ eigenvalues of $U_d U_d^\top$ must be equal to zero. Since σ is as small as possible, the smallest $D - d$ eigenvalues of $\Sigma_{\boldsymbol{x}}$ must be equal to each other and equal to σ^2. To find U_d, let

$$\Sigma_{\boldsymbol{x}} = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Lambda_1 & \mathbf{0} \\ \mathbf{0} & \sigma^2 I_{D-d} \end{bmatrix} \begin{bmatrix} U_1 & U_2 \end{bmatrix}^\top \tag{2.53}$$

be the eigenvalue decomposition of $\Sigma_{\boldsymbol{x}}$, where the columns of U_1 are the top d eigenvectors of $\Sigma_{\boldsymbol{x}}$ and the entries of Λ_1 are the corresponding eigenvalues. Then,

$$U_d U_d^\top = \Sigma_{\boldsymbol{x}} - \sigma^2 I_D = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Lambda_1 - \sigma^2 I_d & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} U_1 & U_2 \end{bmatrix}^\top \tag{2.54}$$

$$= U_1(\Lambda_1 - \sigma^2 I_d) U_1^\top. \tag{2.55}$$

Since both U_d and U_1 are of rank d, all the solutions for U_d must be of the form $U_d = U_1(\Lambda_1 - \sigma^2 I_d)^{1/2} R$, where R is an arbitrary orthogonal matrix. □

2.2.2 *PPCA by Maximum Likelihood*

In practice, we may not know the population mean and covariance, $\boldsymbol{\mu}_{\boldsymbol{x}}$ and $\Sigma_{\boldsymbol{x}}$. Instead, we are given N i.i.d. samples, $\{\boldsymbol{x}_j\}_{j=1}^N$, from which we wish to estimate the PPCA model parameters $\boldsymbol{\mu}$, U_d and σ. In this section, we show that the ML estimates (see Appendix B.1.4) of these parameters can be computed in closed form from the ML estimates of the mean and covariance.

To that end, recall that in the PPCA model $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{x}}, \Sigma_{\boldsymbol{x}})$, where $\boldsymbol{\mu}_{\boldsymbol{x}} = \boldsymbol{\mu}$ and $\Sigma_{\boldsymbol{x}} = U_d U_d^\top + \sigma^2 I_D$. Therefore, the log-likelihood of $\boldsymbol{x}$ is given by

$$\begin{aligned} \mathcal{L} &= \sum_{j=1}^N \log \Big(\frac{1}{(2\pi)^{D/2} \det(\Sigma_{\boldsymbol{x}})^{1/2}} \exp \big(- \frac{(\boldsymbol{x}_j - \boldsymbol{\mu}_{\boldsymbol{x}})^\top \Sigma_{\boldsymbol{x}}^{-1} (\boldsymbol{x}_j - \boldsymbol{\mu}_{\boldsymbol{x}})}{2} \big) \Big) \\ &= -\frac{ND}{2} \log(2\pi) - \frac{N}{2} \log\det(\Sigma_{\boldsymbol{x}}) - \frac{1}{2} \sum_{j=1}^N (\boldsymbol{x}_j - \boldsymbol{\mu})^\top \Sigma_{\boldsymbol{x}}^{-1} (\boldsymbol{x}_j - \boldsymbol{\mu}). \end{aligned} \tag{2.56}$$

We obtain the ML estimate for $\boldsymbol{\mu}$ from the derivatives of $\mathcal{L}$ with respect to $\boldsymbol{\mu}$, to be

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\mu}} = -\sum_{j=1}^N \Sigma_{\boldsymbol{x}}^{-1} (\boldsymbol{x}_j - \boldsymbol{\mu}) = 0 \implies \widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{\mu}}_N \doteq \frac{1}{N} \sum_{j=1}^N \boldsymbol{x}_j. \tag{2.57}$$

After replacing $\widehat{\boldsymbol{\mu}}$ into the log-likelihood, we obtain

$$\mathcal{L} = -\frac{ND}{2} \log(2\pi) - \frac{N}{2} \log\det(\Sigma_{\boldsymbol{x}}) - \frac{N}{2} \operatorname{trace}(\Sigma_{\boldsymbol{x}}^{-1} \widehat{\Sigma}_N), \tag{2.58}$$

where

$$\widehat{\Sigma}_N \doteq \frac{1}{N}\sum_{j=1}^{N}(\boldsymbol{x}_j - \widehat{\boldsymbol{\mu}}_N)(\boldsymbol{x}_j - \widehat{\boldsymbol{\mu}}_N)^\top. \tag{2.59}$$

The answer to the question of whether U_d and σ can be estimated as in Theorem 2.7 after replacing $\Sigma_{\boldsymbol{x}}$ by $\widehat{\Sigma}_N$ is given by the following theorem.

Theorem 2.8 (PPCA by Maximum Likelihood)**.** *The ML estimates for the parameters of the PPCA model $\boldsymbol{\mu}$, U_d and σ can be obtained from the ML estimates of the mean and covariance of the data, $\widehat{\boldsymbol{\mu}}_N$ and $\widehat{\Sigma}_N$, respectively, as*

$$\widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{\mu}}_N, \quad \widehat{U}_d = U_1(\Lambda_1 - \widehat{\sigma}^2 I)^{1/2}R \quad \textit{and} \quad \widehat{\sigma}^2 = \frac{1}{D-d}\sum_{i=d+1}^{D}\lambda_i, \tag{2.60}$$

where U_1 is the matrix with the top d eigenvectors of $\widehat{\Sigma}_N$, Λ_1 is the matrix with the corresponding top d eigenvalues, $R \in \mathbb{R}^{d\times d}$ is an arbitrary orthogonal matrix and λ_i is the i-th largest eigenvalue of $\widehat{\Sigma}_N$.

Proof. We have already shown that $\widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{\mu}}_N$. To find U_d, we need to compute the derivatives of $\mathcal{L}$ with respect to U_d. It follows from Exercise A.4 that $\frac{\partial}{\partial X}\log(|\det(X)|) = (X^{-1})^\top$, $\frac{\partial}{\partial X}\operatorname{trace}(AX^{-1}B) = -(X^{-1}BAX^{-1})^\top$ and $\frac{\partial}{\partial X}\operatorname{trace}(XBX^\top) = XB^\top + XB$. Therefore,

$$\frac{\partial\mathcal{L}}{\partial U_d} = -N\Sigma_{\boldsymbol{x}}^{-1}U_d + N\Sigma_{\boldsymbol{x}}^{-1}\widehat{\Sigma}_N\Sigma_{\boldsymbol{x}}^{-1}U_d = 0 \implies \widehat{\Sigma}_N\Sigma_{\boldsymbol{x}}^{-1}U_d = U_d. \tag{2.61}$$

One possible solution is $U_d = 0$, which leads to a minimum of the log-likelihood and violates our assumption that U_d should be full rank. Another possible solution is $\Sigma_{\boldsymbol{x}} = \widehat{\Sigma}_N$, where the covariance model is exact. This corresponds to the case discussed in the previous section, after replacing $\Sigma_{\boldsymbol{x}}$ by $\widehat{\Sigma}_N$. Thus, the model parameters can be computed as in Theorem 2.7 as equation (2.60) reduces to equation (2.52). A third solution is obtained when $U_d \neq 0$ and $\Sigma_{\boldsymbol{x}} \neq \widehat{\Sigma}_N$. In this case, let $U_d = W\Gamma V^\top$ be the compact SVD of U_d, where $W \in \mathbb{R}^{D\times d}$ is a matrix with orthonormal columns, $\Gamma \in \mathbb{R}^{d\times d}$ is an invertible diagonal matrix, and $V \in \mathbb{R}^{d\times d}$ is an orthogonal matrix. Then

$$\Sigma_{\boldsymbol{x}} = W\Gamma^2W^\top + \sigma^2 I_D = W(\Gamma^2 + \sigma^2 I_d)W^\top + \sigma^2 W^\perp W^{\perp\top}, \tag{2.62}$$

where $W^\perp \in \mathbb{R}^{D\times(D-d)}$ is an orthonormal matrix such that $W^\top W^\perp = \mathbf{0}$. Thus,

$$\widehat{\Sigma}_N\Sigma_{\boldsymbol{x}}^{-1}U_d = \widehat{\Sigma}_N(W(\Gamma^2 + \sigma^2 I_d)^{-1}W^\top + \sigma^{-2}W^\perp W^{\perp\top})W\Gamma V^\top \tag{2.63}$$

$$= \widehat{\Sigma}_N W(\Gamma^2 + \sigma^2 I_d)^{-1}\Gamma V^\top = W\Gamma V^\top \tag{2.64}$$

and

$$\widehat{\Sigma}_N W = W(\Gamma^2 + \sigma^2 I_d). \tag{2.65}$$

Letting $W = [w_1, \ldots, w_d]$ and $\Gamma = \operatorname{diag}\{\gamma_1, \ldots, \gamma_d\}$, we obtain

$$\widehat{\Sigma}_N w_i = (\gamma_i^2 + \sigma^2)w_i \quad \forall i = 1, \ldots, d. \tag{2.66}$$

Hence, W is a matrix containing d eigenvectors of $\widehat{\Sigma}_N$ with corresponding eigenvalues $\gamma_i^2 + \sigma^2$. Let $\widehat{\Sigma}_N = U\Lambda U^\top = [U_1, U_2]\text{diag}\{\Lambda_1, \Lambda_2\}[U_1, U_2]^\top$ be the eigenvalue decomposition of $\widehat{\Sigma}_N$, where we partition U and Λ so that the d chosen eigenvectors and eigenvalues are in U_1 and Λ_1, respectively. Then, all optimal solutions for U_d are of the form

$$U_d = W\Gamma V^\top = U_1(\Lambda_1 - \sigma^2 I_d)^{1/2} V^\top. \tag{2.67}$$

To determine σ, we replace the solution for U_d into the likelihood in (2.58). Noticing that

$$\begin{aligned}\det(\Sigma_{\boldsymbol{x}}) &= \det\left(U_d U_d^\top + \sigma^2 I_D\right) &(2.68)\\ &= \det\left(U_1(\Lambda_1 - \sigma^2 I_d)U_1^\top + \sigma^2(U_1U_1^\top + U_2U_2^\top)\right) &(2.69)\\ &= \det(U_1\Lambda_1U_1^\top + \sigma^2 U_2U_2^\top) = \det(\Lambda_1)\sigma^{2(D-d)} &(2.70)\end{aligned}$$

and that

$$\begin{aligned}\text{trace}(\Sigma_{\boldsymbol{x}}^{-1}\widehat{\Sigma}_N) &= \text{trace}\,(U_1\Lambda_1^{-1}U_1^\top + \sigma^{-2}U_2U_2^\top)(U_1\Lambda_1U_1^\top + U_2\Lambda_2U_2^\top) &(2.71)\\ &= \text{trace}\,U_1U_1^\top + \sigma^{-2}U_2\Lambda_2U_2^\top = d + \sigma^{-2}\,\text{trace}\,\Lambda_2, &(2.72)\end{aligned}$$

we obtain

$$\mathcal{L} = -\frac{N}{2}\left(D\log(2\pi) + \log\det(\Lambda_1) + (D-d)\log\sigma^2 + d + \sigma^{-2}\,\text{trace}(\Lambda_2)\right). \tag{2.73}$$

The condition for an extremum in σ^2 is given by

$$\frac{\partial\mathcal{L}}{\partial\sigma^2} = -\frac{N}{2}\left(\frac{D-d}{\sigma^2} - \frac{\text{trace}(\Lambda_2)}{\sigma^4}\right) = 0 \implies \sigma^2 = \frac{\text{trace}(\Lambda_2)}{D-d}. \tag{2.74}$$

Therefore, σ^2 is the average of the discarded eigenvalues of $\widehat{\Sigma}_N$.

To determine which d eigenvectors and eigenvalues of $\widehat{\Sigma}_N$ should be discarded, notice that $\det(\Lambda_1) = \frac{\det(\Lambda)}{\det(\Lambda_2)}$. Hence, after replacing the optimal σ^2 into $\mathcal{L}$, we can see that the maximization of $\mathcal{L}$ is equivalent to the minimization of

$$\mathcal{M} = \log\left(\frac{\sum_{i=d+1}^D \lambda_{\pi[i]}}{D-d}\right) - \frac{\sum_{i=d+1}^D \log\lambda_{\pi[i]}}{D-d}, \tag{2.75}$$

where $\lambda_{\pi[1]}, \dots, \lambda_{\pi[d]}$ are the chosen eigenvalues and $\lambda_{\pi[d+1]}, \dots, \lambda_{\pi[D]}$ are the discarded ones. Since $\log$ is a concave downwards function, by Jensen's inequality, $\mathcal{M}$ is nonnegative. Since the $\log$ function is concave downwards, the reader can verify (see Exercise 2.10) that $\mathcal{M}$ is minimized when the discarded eigenvalues are contiguous within the spectrum of the ordered eigenvalues of $\widehat{\Sigma}_N$. Further, since the chosen eigenvalues must be such that $\lambda_{\pi[i]} \geq \sigma^2$ for $i = 1, \dots, d$, the discarded eigenvalues must be the $D - d$ smallest eigenvalues. Indeed if not, $\lambda_{\min} = \min\limits_{i=1,\dots,D} \lambda_i$ would be one of the chosen eigenvalues and we would have $\lambda_{\min} < \sigma^2$, which would be a contradiction to equation (2.66). Therefore, the optimal solutions for U_d and σ are given by (2.60). □

2.3 Model Selection for Principal Component Analysis

One of the main goals of both PCA and PPCA is to reduce the data to a small number of principal components that capture as much information about the data as possible. So far, we have assumed that the number of principal components, d, or the dimension of the subspace S, is known. In practice, however, we may not know the intrinsic dimension of the data. In this section, we review a few methods (several of them heuristic) for estimating the number of principal components.

2.3.1 Model Selection by Information Theoretic Criteria

Let $X = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N] \in \mathbb{R}^{D \times N}$ be the mean-subtracted data matrix. When the data points are noise free, they lie exactly in a subspace of dimension d. Hence, we can estimate d as the rank of X, i.e., $d = \operatorname{rank}(X)$. However, when the data are contaminated by noise, the matrix X will be full rank in general, hence we cannot use its rank to estimate d. Nonetheless, notice that the SVD of the noisy data matrix X gives a solution to PCA not only for a particular dimension of the subspace, d, but also for all $d = 1, 2, \ldots, D$. This has an important side-benefit: If the dimension of the subspace S is *not* known or specified a priori, rather than optimizing for both d and S simultaneously, we can easily look at the entire spectrum of solutions for different values of d to decide on the "best" estimate $\widehat{d}$ for the dimension of the subspace d given the data X.

One possible criterion is to chose d as the dimension that minimizes the least-squares error between the given data X and its projection $\widehat{X}^d = [\widehat{\boldsymbol{x}}_1^d, \widehat{\boldsymbol{x}}_2^d, \ldots, \widehat{\boldsymbol{x}}_N^d]$ onto the subspace S of dimension d. As shown in the proof of Theorem 2.3, the least-squares error is given by the sum of the squares of the remaining singular values of X, i.e.,

$$J(d) \doteq \|X - \widehat{X}^d\|_F^2 = \sum_{j=1}^{N} \|\boldsymbol{x}_j - \widehat{\boldsymbol{x}}_j^d\|^2 = \sum_{i=d+1}^{D} \sigma_i^2. \tag{2.76}$$

However, this is not a good criterion, because $J(d)$ is a non-increasing function of d. In fact, the best solution is obtained when $d = \operatorname{rank}(X)$, because $J(d) = 0$.

The problem of determining the optimal dimension $\widehat{d}$ is in fact a "model selection" problem. As we discussed in the introduction of the book, the conventional wisdom is to strike a good balance between the *complexity* of the chosen model and the *fidelity* of the data to the model. The dimension d of the subspace S is a natural measure of model complexity, while the least-squares error $\|X - \widehat{X}^d\|_F^2 = \sum_{i=d+1}^{D} \sigma_i^2$ or its leading term σ_{d+1}^2 are natural measures of the data fidelity.

Perhaps the simplest model selection criterion is to minimize the complexity subject to a bound on the fidelity. For example, we can choose d as the smallest

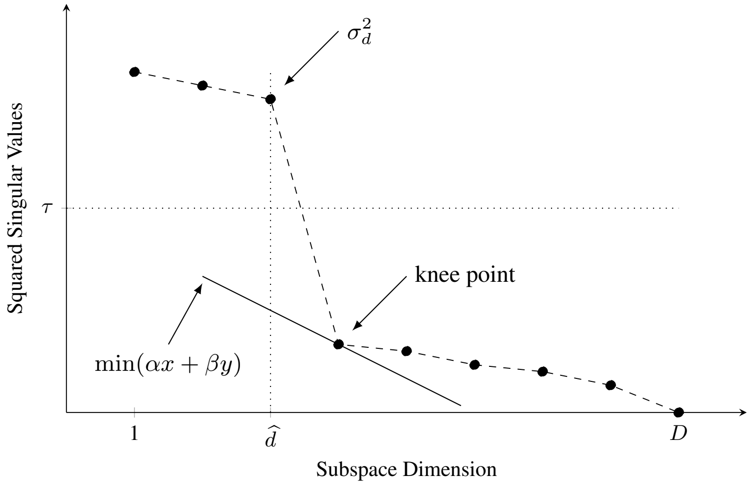

Figure 2.2. Singular value as a function of the dimension of the subspace.

number such that the fidelity is less than a threshold $\tau > 0$, i.e.,

$$\widehat{d} = \min_d \left\{ d : \sum_{i=d+1}^{D} \sigma_i^2 < \tau \right\} \quad \text{or} \quad \widehat{d} = \min_d \left\{ d : \sigma_{d+1}^2 < \tau \right\}. \tag{2.77}$$

The second criterion is illustrated in Figure 2.2. However, it is very hard to choose an appropriate τ, because the singular values of X are not invariant with respect to transformations of the data, such as scaling. One possible solution is to normalize the singular values by $\|X\|_F^2 = \sum_{i=1}^D \sigma_i^2$ and estimate d as

$$\widehat{d} = \min_d \left\{ d : \frac{\sum_{i=d+1}^{D} \sigma_i^2}{\sum_{i=1}^{D} \sigma_i^2} < \tau \right\} \quad \text{or} \quad \widehat{d} = \min_d \left\{ d : \frac{\sigma_{d+1}^2}{\sum_{i=1}^{D} \sigma_i^2} < \tau \right\}. \tag{2.78}$$

The first criterion is widely used, because it has an intuitive interpretation: the number of principal components is chosen as the smallest number such that the fraction of information being discarded is less than a threshold τ. Typical values for τ are in the range 10-20%.

Yet another model selection criterion seeks a balance between d and σ_{d+1}^2 by minimizing an objective function of the form:

$$\hat{d} = \arg\min \tilde{J}(d) \doteq \alpha \cdot \sigma_{d+1}^2 + \beta \cdot d \tag{2.79}$$

for some proper weights $\alpha, \beta > 0$. In general, the ordered squared singular values of the data matrix X versus the dimension d of the subspace resemble a plot similar to that shown in Figure 2.2. In the statistics literature, this is known as the "scree graph," which was discussed and named by [Cattell, 1966]. We should see a significant drop in the singular values right after the "correct" dimension $\widehat{d}$, which is sometimes called the "knee" or "elbow" point of the plot. Such a point is a stable minimum as it optimizes the above objective function (2.79) for a range of values for α and β.

A more principled approach to finding the optimal dimension of the subspace, $\widehat{d}$, is to use some of the model selection criteria described in Appendix B. Such criteria rely on a different choice of the model complexity term and provide an automatic way of choosing the parameters α and β. Specifically, the complexity of the model is measured by the number of parameters needed to describe the subspace. This count is made by using the so-called Grassmann coordinates, which give the dimension of the parameter space for a d-dimensional subspace in $\mathbb{R}^D$ to be $Dd-d^2$.[4] With a model parameter space of dimension $Dd-d^2$ and a Gaussian noise model with known variance σ^2, the Bayesian information criterion (BIC) is equivalent to minimizing

$$\mathrm{BIC}(d) \doteq \sum_{i=d+1}^{D} \sigma_i^2 + (\log N)(Dd - d^2)\sigma^2, \tag{2.80}$$

while the Akaike information criterion (AIC) minimizes

$$\mathrm{AIC}(d) \doteq \sum_{i=d+1}^{D} \sigma_i^2 + 2(Dd - d^2)\sigma^2. \tag{2.81}$$

More recently, a geometric version of the Akaike information criterion has been proposed by [Kanatani, 2003]. The Geometric AIC minimizes

$$\text{G-AIC}(d) \doteq \sum_{i=d+1}^{D} \sigma_i^2 + 2(Dd - d^2 + Nd)\sigma^2, \tag{2.82}$$

where the extra term Nd accounts for the number of coordinates needed to represent (the closest projection of) the given N data points in the estimated d-dimensional subspace. From an information-theoretic viewpoint, the additional Nd coordinates are necessary if we are interested in encoding not only the model but also the data themselves. This is often the case when we use PCA for purposes such as data compression and dimension reduction. The quantity $\frac{(Dd-d^2+Nd)}{N}$ is closely related to the so-called "effective dimension" of the data set defined in [Huang et al., 2004], which can be generalized to multiple subspaces.

In some sense, all the above criteria can be loosely referred to as *information-theoretic* model selection criteria, in the sense that most of these criteria can be interpreted as variations to minimizing the optimal code length for both the model and the data with respect to certain class of distributions and coding

[4] $Dd - d^2$ is the dimension of the Grassmann manifold of d-dimensional subspaces in $\mathbb{R}^D$. To specify a subspace, one can use the so-called Grassmann coordinates which need exactly $Dd - d^2$ entries: starting with a $D \times d$ matrix whose columns form a basis for the subspace, perform column-reduction so that the first $d \times d$ block is the identity matrix. Then, one only needs to give the remaining $(D - d) \times d$ entries to specify the subspace.

schemes [Hansen and Yu, 2001].[5] There are many other methods for determining the number of principal components. The interested reader may find more references in [Jolliffe, 1986b].

2.3.2 *Model Selection by Rank Minimization*

In this section, we present an alternative view of model selection based on the rank minimization approach to PCA introduced in Section 2.1.3. In this approach, the PCA problem is posed as one of finding a rank-d matrix A that best approximates the mean substracted data matrix X, i.e.,

$$\min_A \ \|X - A\|_F^2 \ \text{ s.t. } \ \operatorname{rank}(A) = d. \tag{2.83}$$

Although this problem is non-convex due to the rank constraint, as we showed in Section 2.1.3, its optimal solution can be computed in closed form as

$$A = U\mathcal{H}_{\sigma_{d+1}}(\Sigma)V^\top, \tag{2.84}$$

where $X = U\Sigma V^\top$ is the SVD of X, σ_k is the k-th singular value of X, and $\mathcal{H}_\varepsilon(x)$ is the *hard thresholding operator*:

$$\mathcal{H}_\varepsilon(x) = \begin{cases} x & |x| > \varepsilon \\ 0 & \text{else} \end{cases}. \tag{2.85}$$

However, this closed solution requires d to be known.

When d is unknown, the problem of finding a low-rank approximation can be formulated as

$$\min_A \quad \|X - A\|_F^2 + \tau \operatorname{rank}(A), \tag{2.86}$$

where $\tau > 0$ is a parameter. Since the optimal solution of (2.83) for a fixed rank $d = \operatorname{rank}(A)$ is $A = U\mathcal{H}_{\sigma_{d+1}}(\Sigma)V^\top$, the problem in (2.86) is equivalent to

$$\min_d \ \sum_{k>r} \sigma_k^2 + \tau d. \tag{2.87}$$

The optimal solution is the smallest d such that $\sigma_{d+1}^2 \leq \tau$. Notice that this model selection criteria is the same as that in (2.77). Therefore, the optimization problem in (2.86) provides a justification for the criteria in (2.77). Under this criteria, and with the notation introduced in this section, the optimal A is given by

$$A = U\mathcal{H}_{\sqrt{\tau}}(\Sigma)V^\top. \tag{2.88}$$

[5] Even if one chooses to compare models by their algorithmic complexity, such as the minimum message length (MML) criterion [Wallace and Boulton, 1968] (an extension of the Kolmogrov complexity to model selection), a strong connection with the above information-theoretic criteria, such as MDL, can be readily established via Shannon's optimal coding theory (see [Wallace and Dowe, 1999]).

Therefore, the optimal A can still be computed in closed form from the SVD of X, in spite of the fact that the optimization problem in (2.86) is non convex.

Most rank minimization problems are, however, NP hard and cannot be solved as easily as the one in (2.86). This has motivated the development of convex relaxations, which lead to more efficient solutions. A commonly used relaxation (see e.g., [Cai et al., 2008, Recht et al., 2010]) is to replace the rank of A by its nuclear norm $\|A\|_* = \sum \sigma_k(A)$, i.e., the sum of its singular values. As it turns out, this relaxation leads to a slightly different model selection criteria for PCA. More specifically, the relaxation of (2.86) (modulo the $1/2$ factor) is given by

$$\min_A \quad \frac{1}{2}\|X - A\|_F^2 + \tau\|A\|_*. \tag{2.89}$$

The sub-gradient of this function w.r.t. A is given by $A - X + \partial\|A\|_*$, where $\partial\|A\|_*$ is the sub-gradient of the nuclear norm of A (see Exercise 3.7). Therefore, as shown in [Cai et al., 2008] (see also Exercise 3.8), the optimal solution for A is given by

$$A = \mathcal{D}_\tau(X) = U\mathcal{S}_\tau(\Sigma)V^\top, \tag{2.90}$$

where $\mathcal{D}_\varepsilon$ is the *singular value thresholding operator* and $\mathcal{S}_\varepsilon$ is the *soft thresholding operator*, which is defined as

$$\mathcal{S}_\varepsilon(x) = \text{sign}(x)\max(|x| - \varepsilon, 0) = \begin{cases} x - \varepsilon & x > \varepsilon \\ x + \varepsilon & x < -\varepsilon \\ 0 & \text{else} \end{cases}. \tag{2.91}$$

Notice that the latter solution does not coincide with the one given by PCA, which performs hard-thresholding of the singular values of X without shrinking them by τ. However, the model selection criteria is the same: choose d as the largest integer such that $\sigma_{d+1}^2 > \tau$.

2.3.3 Model Selection by Asymptotic Mean Square Error

From the above two sections, we see that by following different model selection criteria or objectives, we essentially have three different types of estimators $\hat{X}$ for a low-rank matrix X_0 from its noisy measurements: $X = X_0 + \sigma Z$. Let the SVD of X to be $X = U\Sigma V^\top$, the three estimators are of the following forms, respectively:

1. If the rank of d is known, the optimal estimate $\hat{X}$ subject to $\mathrm{rank}(\hat{X}) = d$, is the *truncated SVD* solution:

   $$\hat{X}_1 = U\mathcal{T}_d(\Sigma)V^\top.$$

   Or if the rank d is not known and one uses one of the information-theoretic criteria given in Section 2.3.1 to estimate the dimension $\hat{d}$. Then we only have to replace the d in the above solution with the estimated $\hat{d}$.

2. If we try to balance the mean squared error and the dimension as in equation (2.86), the optimal estimate is given by the *SVD hard thresholding*:

   $$\hat{X}_2 = U\mathcal{H}_\tau(\Sigma)V^\top.$$

   for some threshold $\tau > 0$.

3. If we try to balance the mean squared error and the nuclear norm as in equation (2.89), the optimal estimate is given by the *SVD soft thresholding*:

   $$\hat{X}_3 = U\mathcal{S}_\tau(\Sigma)V^\top.$$

   for some threshold $\tau > 0$.

Naturally, this may lead to certain degree of confusion for practitioners: Which estimate is "the best"? What is the optimal threshold τ^* to use in case we need to threshold the singular values? Which thresholding is better, hard or soft? The short answer to these questions is that none of the above estimators is always better than others, as they are all optimal in their own way under certain conditions.

However, if we all agree on a common objective based on a common noise model, it might be meaningful and even insightful to examine which estimator is better than others. One such setting was recently proposed by [Donoho and Gavish, 2013]. That is to study the different estimators in terms of their Mean Square Errors (MSE) in an asymptotic setting as the size of the matrix becomes large:

$$\text{AMSE} = \lim_{N\to\infty} \|\hat{X} - X_0\|_F^2.$$

As it turns out, this would allow us to find clear answers to the above questions with some additional useful findings.

For simplicity, we first assume the matrix X is a square matrix of size $N = D$. In the asymptotic setting (as $N \to \infty$), we assume the following noise model:

$$X = X_0 + \sigma Z \tag{2.92}$$

where Z is a matrix whose entries are i.i.d. drawn from a probability (say Gaussian) distribution with zero mean and variance $1/\sqrt{N}$. It is easy to see that the noise level in the singular values of X is σ. Among all estimates of X_0 obtained by a hard thresholding of the singular values of X, we are interested in finding the one that minimizes the asymptotic mean square error. The work of [Donoho and Gavish, 2013] gives the following answer to this question.

Proposition 2.9 (Optimal Hard Thresholding for Minimizing AMSE)**.** *Given a low-rank matrix $X_0 \in \mathbb{R}^{D\times N}$ and noisy measurements $X = X_0 + \sigma Z$ with Z zero mean and variance $1/\sqrt{N}$. If the matrix is square, i.e. $D = N$, the optimal hard threshold estimate $\hat{X} = U\mathcal{H}_{\tau^*}(\Sigma)V^\top$ that minimizes the asymptotic mean square error $\|\hat{X} - X_0\|_F^2$ is given by*

$$\tau^* = 4/\sqrt{3}\sigma \approx 2.309\sigma. \tag{2.93}$$

In the more general case of a non square matrix with $D/N \to \beta$, the optimal threshold is given by

$$\tau^*(\beta) = \sigma\sqrt{2(\beta+1) + \frac{8\beta}{(\beta+1)+\sqrt{\beta^2+14\beta+1}}}.$$

The proof of this statement is beyond the scope of this book. But that does not prevent us from discussing and understanding its implications in our context.

In can be shown that under the same noise model, the distribution of the singular values of the matrix $X = X_0 + \sigma Z$ form a quarter-circle bulk, whose radius lies approximately at $(1+\sqrt{\beta})\sigma$. This is the place where we would normally expect to see a "knee point" in the distribution of singular values (as shown in Figure 2.2). The information-theoretic criteria or the rank-minimization objectives are most likely to pick this value to threshold the singular values. For a square matrix, this gives the threshold $\tau = 2\sigma$, which is close but not quite at the optimal value 2.309σ. As shown in the work of [Donoho and Gavish, 2013], this minor difference in the choice of the threshold can result in a $5/3$-fold increase in AMSE.

Interestingly, even if we know the correct rank d of the matrix X_0 and take the truncated SVD solution $\hat{X} = U\mathcal{T}_d(\Sigma)V^\top$, the resulting AMSE is also $5/3$-fold larger than that of the optimal hard thresholding solution given above. In general, soft thresholding does not work as well as hard thresholding in the high signal-to-noise ratio regime, and the AMSE for the optimal soft thresholding solution $\hat{X}_3 = U\mathcal{S}_{\tau^*}(\Sigma)V^\top$ is twice as large as that of hard thresholding. In fact, even if one is allowed to use any singular value shrinkage function instead of merely a hard or soft thresholding (see the work of [Shabalin and Nobel, 2010] for more details), compared to the above optimal hard thresholding solution (2.93), one can at best reduce the AMSE by another $1/3$.

2.4 Bibliographic Notes

As a matrix decomposition tool, SVD was initially developed independently from PCA in the numerical linear algebra literature, also known as the Eckart and Young decomposition [Eckart and Young, 1936, Hubert et al., 2000]. The result regarding the least-squares optimality of SVD given in Theorem 2.3 can be traced back to [Householder and Young, 1938, Gabriel, 1978]. While principal components were initially defined exclusively in a statistical sense [Pearson, 1901, Hotelling, 1933], one can show that the algebraic solution given by SVD gives asymptotically unbiased estimates of the true parameters in the case of Gaussian distributions. A more detailed analysis of the statistical properties of PCA can be found in [Jolliffe, 2002].

Note that PCA only infers the principal subspace (or components), but not a probabilistic distribution of the data in the subspace. Probabilistic PCA was developed to infer an explicit probabilistic distribution from the data [Tipping and Bishop, 1999b]. The data is assumed to be independent samples drawn from an unknown distribution, and the problem becomes one of identifying the subspace and the parameters of the distribution in a maximum-likelihood or a maximum-a-posteriori sense. When the underlying noise distribution is Gaussian, the geometric and probabilistic interpretations of PCA coincide [Collins et al., 2001]. However, when the underlying distribution is non Gaussian, the optimal solution to PPCA may no longer be linear. For example, in [Collins et al., 2001] PCA is generalized to arbitrary distributions in the exponential family.

2.5 Exercises

Exercise 2.1 (Properties of Symmetric Matrices). Let $S \in \mathbb{R}^{n\times n}$ be a real symmetric matrix. Prove the following:

1. All the eigenvalues of S are real, i.e., $\sigma(S) \subset \mathbb{R}$.
2. Let (λ, v) be an eigenvalue-eigenvector pair. If $\lambda_i \neq \lambda_j$, then $v_i \perp v_j$; i.e., eigenvectors corresponding to distinct eigenvalues are orthogonal.
3. There always exist n orthonormal eigenvectors of S, which form a basis of $\mathbb{R}^n$.
4. S is positive definite (positive semidefinite) if and only if all of its eigenvalues are positive (non-negative), i.e., $S \succ 0$ ($S \succeq 0$), iff $\forall i = 1, 2, \ldots, n$, $\lambda_i > 0$ ($\lambda_i \geq 0$).
5. If $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ are the sorted eigenvalues of S, then $\max_{\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^\top S \boldsymbol{x} = \lambda_1$ and $\min_{\|x\|_2=1} \boldsymbol{x}^\top S \boldsymbol{x} = \lambda_n$.

Exercise 2.2 (Pseudo-inverse of a Matrix).

1. Let $A = U_r \Sigma_r V_r^\top$ be the compact SVD of a matrix A of rank r. Show that the pseudo-inverse of A is given by $A^\dagger = V_r \Sigma_r^{-1} U_r^\top$.
2. Consider the linear system of equations $A\boldsymbol{x} = \boldsymbol{b}$, where the matrix $A \in \mathbb{R}^{m\times n}$ is of rank $r = \operatorname{rank}(A) = \min\{m, n\}$. Show that the solution $\boldsymbol{x}^*$ that minimizes $\|A\boldsymbol{x} - \boldsymbol{b}\|_2^2$ is given by $\boldsymbol{x}^* = A^\dagger \boldsymbol{b}$, where $A^\dagger$ is the pseudo-inverse of A.

Exercise 2.3 (Statistical PCA for Non-Zero Mean Random Variables) Let $\boldsymbol{x} \in \mathbb{R}^D$ be a random vector. Let $\boldsymbol{\mu_x} = \mathbb{E}(\boldsymbol{x}) \in \mathbb{R}^D$ and $\Sigma_{\boldsymbol{x}} = \mathbb{E}(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^\top \in \mathbb{R}^{D\times D}$ be, respectively, the mean and the covariance of $\boldsymbol{x}$. Define the principal components of $\boldsymbol{x}$ as the random variables $y_i = \boldsymbol{u}_i^\top \boldsymbol{x} + a_i \in \mathbb{R}$, $i = 1, \ldots, d \leq D$, where $\boldsymbol{u}_i \in \mathbb{R}^D$ is a unit norm vector, $a_i \in \mathbb{R}$, and $\{y_i\}_{i=1}^n$ are zero mean, uncorrelated random variables whose variances are such that Var$(y_1) \geq$ Var$(y_2) \geq \cdots \geq$ Var(y_d). Assuming that the eigenvalues of $\Sigma_{\boldsymbol{x}}$ are different from each other, show that

1. $a_i = -\boldsymbol{u}_i^\top \boldsymbol{\mu_x}$, $i = 1, \ldots, d$.
2. $\boldsymbol{u}_1$ is the eigenvector of $\Sigma_{\boldsymbol{x}}$ corresponding to its largest eigenvalue.
3. $\boldsymbol{u}_2^\top \boldsymbol{u}_1 = 0$ and $\boldsymbol{u}_2$ is the eigenvector of Σ corresponding to its second largest eigenvalue.

4. $\boldsymbol{u}_i^\top \boldsymbol{u}_j = 0$ for all $i \neq j$ and $\boldsymbol{u}_i$ is the eigenvector of $\Sigma_{\boldsymbol{x}}$ corresponding to its i-th largest eigenvalue.

Exercise 2.4 (Properties of PCA). Let $\boldsymbol{x} \in \mathbb{R}^D$ be a random vector with covariance matrix $\Sigma_{\boldsymbol{x}} \in \mathbb{R}^{D\times D}$. Consider a linear transformation of $\boldsymbol{x}$:

$$\boldsymbol{y} = W^\top \boldsymbol{x}, \tag{2.94}$$

where $\boldsymbol{y} \in \mathbb{R}^d$ and $W \in \mathbb{R}^{D\times d}$ has orthonormal columns. Let $\Sigma_{\boldsymbol{y}} = W^\top \Sigma_{\boldsymbol{x}} W$ be the covariance matrix for $\boldsymbol{y}$. Show that

1. The trace of $\Sigma_{\boldsymbol{y}}$ is maximized by $W = U_d$, where U_d consists of the first d unit eigenvectors of $\Sigma_{\boldsymbol{x}}$.
2. The trace of $\Sigma_{\boldsymbol{y}}$ is minimized by $W = \tilde{U}_d$, where $\tilde{U}_d$ consists of the last d unit eigenvectors of Σ.

Exercise 2.5 (Subspace Angles). Given two d-dimensional subspaces S_1 and S_2 in $\mathbb{R}^D$, define the largest subspace angle θ_1 between S_1 and S_2 to be the largest possible sharp angle ($< 90°$) formed by any two vectors $\boldsymbol{u}_1, \boldsymbol{u}_2 \in (S_1 \cap S_2)^\perp$ with $\boldsymbol{u}_1 \in S_1$ and $\boldsymbol{u}_2 \in S_2$ respectively. Let $U_1 \in \mathbb{R}^{D\times d}$ be an orthogonal matrix whose columns form a basis for S_1 and similarly U_2 for S_2. Then show that if σ_1 is the smallest non-zero singular value of the matrix $W = U_1^\top U_2$, then we have

$$\cos(\theta_1) = \sigma_1. \tag{2.95}$$

Similarly, one can define the rest of the subspace angles as $\cos(\theta_i) = \sigma_i, i = 2, \dots, d$ from the rest of the singular values of W.

Hint: Following the derivation of statistical PCA, find first the smallest angle (largest cosine = largest variance) and then find the second smallest angle all the way to the largest angle (smallest variance). As your proceed, the vectors that achieve the second smallest angle need to be chosen to be perpendicular to the vectors that achieve the smallest angle and so forth, as we did in statistical PCA. Also, let $\boldsymbol{u}_1 = U_1 \boldsymbol{c}_1$ and $\boldsymbol{u}_2 = U_2 \boldsymbol{c}_2$. Show that you need to optimize $\cos(\theta) = \boldsymbol{c}_1^\top U_1^\top U_2 \boldsymbol{c}_2$ subject to $\|\boldsymbol{c}_1\| = \|\boldsymbol{c}_2\| = 1$. Show (using Lagrange multipliers) that a necessary condition for optimality is

$$\begin{bmatrix} 0 & U_1^\top U_2 \\ U_2^\top U_1 & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{c}_1 \\ \boldsymbol{c}_2 \end{bmatrix} = \lambda \begin{bmatrix} \boldsymbol{c}_1 \\ \boldsymbol{c}_2 \end{bmatrix}. \tag{2.96}$$

Deduce from here that $\sigma = \lambda^2$ is a singular value of $U_1^\top U_2$ with $\boldsymbol{c}_2$ as singular vector.

Exercise 2.6 (Fixed-Rank Approximation of a Matrix). Let $A = U\Sigma V^\top$ be the SVD of A. Let $B = U\Sigma_p V^\top$, where Σ_p denotes the matrix obtained from Σ by setting to zero its elements on the diagonal after the p-th entry. Show that $\|A - B\|_F^2 = \sigma_{p+1}^2 + \cdots + \sigma_r^2$, where $\|\cdot\|_F$ indicates the Frobenius norm. Furthermore, show that such a norm is the minimum achievable over all matrices $B \in \mathbb{R}^{m\times n}$ of rank p, i.e.,

$$\min_{B:\mathrm{rank}(B)=p} \|A - B\|_F^2 = \sigma_{p+1}^2 + \cdots + \sigma_r^2. \tag{2.97}$$

Exercise 2.7 (Identification of Auto-Regressive Exogeneous (ARX) Systems). A popular model that is often used to analyze a time series $\{y_t\}_{t\in\mathbb{Z}}$ is the linear auto-regressive model:

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + \cdots + a_n y_{t-n} + \varepsilon_t, \quad \forall t, y_t \in \mathbb{R}, \tag{2.98}$$

where $\varepsilon_t \in \mathbb{R}$ models the modeling error or noise and it is often assumed to be a white-noise random process. Now suppose that you are given the values of y_t for a sufficiently long period of time.

1. Show that in the noise free case, i.e. $\varepsilon_t \equiv 0$, regardless of the initial conditions, the vectors $\boldsymbol{x}_t = [y_t, y_{t-1}, \ldots, y_{t-n}]^\top$ for all t lie on an n-dimensional hyperplane in $\mathbb{R}^{n+1}$. What is the normal vector to this hyperplane?
2. Now consider the case with noise. Describe how you may use PCA to identify the unknown model parameters $(a_1, a_2, \ldots, a_n)$?

Exercise 2.8 (Basis for an Image). Given a gray-level image $\boldsymbol{I}$, consider all of its $b \times b$ blocks, denoted as $\{B_i \in \mathbb{R}^{b\times b}\}$. We would like to approximate each block as a superposition of d base blocks, say $\{\hat{B}_j \in \mathbb{R}^{b\times b}\}_{j=1}^d$. That is,

$$B_i = \sum_{j=1}^{d} a_{ij}\hat{B}_j + E_i, \tag{2.99}$$

where $E_i \in \mathbb{R}^{b\times b}$ is the possible residual from the approximation. Describe how you can use PCA to identify an optimal set of d base blocks so that the residual is minimized?[6]

Exercise 2.9 (Ranking of Webpages). PCA is actually used to rank webpages on the Internet by many popular search engines. One way to see this is to view the Internet as a directed graph $G = (V, E)$, where every webpage, denoted as p_i, is a node in V, and every hyperlink from p_i to p_j, denoted as e_{ij}, is a directed edge in E. We can assign each webpage p_i an "authority" score x_i and a "hub" score y_i. The "authority" score x_i is a scaled sum of the "hub" scores of other webpages pointing to webpage p_i. The "hub" score is the scaled sum of the "authority" scores of other webpages that webpage p_i is pointing out to. Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be the vector of authority scores and hub scores, respectively. Also, let A be the adjacent matrix of the graph G, i.e., $A_{ij} = 1$ if $e_{ij} \in E$ and $A_{ij} = 0$ otherwise and consider the following algorithm:

Algorithm 2.1 (Ranking webpages)

```
Choose a random vector $\boldsymbol{x}$, and repeat the following two steps
  1. $\boldsymbol{y}' \leftarrow A\boldsymbol{x}, \boldsymbol{y} \leftarrow \frac{\boldsymbol{y}'}{\|\boldsymbol{y}'\|}$
  2. $\boldsymbol{x}' \leftarrow A^\top\boldsymbol{y}, \boldsymbol{x} \leftarrow \frac{\boldsymbol{x}'}{\|\boldsymbol{x}'\|}$
```

Answer the following questions.

1. Given the definitions of hubs and authorities, justify the algorithm.
2. Show that unit-norm eigenvectors of $AA^\top$ (for $\boldsymbol{y}$) and $A^\top A$ (for $\boldsymbol{x}$) give fixed points of the algorithm.

[6] In Section 1.2.1, we have seen an example in which a similar process can be applied to an ensemble of face images, where the first $d = 3$ principal components are computed for further classification. In the computer vision literature, the corresponding base images are called "eigenfaces."

3. Show that, in general, $\boldsymbol{y}$ and $\boldsymbol{x}$ converge to the unit-norm eigenvectors associated with the maximum eigenvalue of $AA^\top$ and $A^\top A$, respectively. Explain why not any other eigenvector and why the normalization steps in the algorithm are necessary.

4. Explain how $\boldsymbol{y}$ and $\boldsymbol{x}$ can be computed from the singular value decomposition of A. Under what circumstances would the given algorithm be preferable to using the SVD?

In the literature, this is known as the *Hypertext Induced Topic Selection* (HITS) algorithm [Kleinberg, 1999, Ding et al., 2004]. The same algorithm can also be used to rank any competitive sports such as football teams and chess players.

Exercise 2.10 Use the concavity of the log function to prove that the $\mathcal{M}$ in equation (2.75) is minimized by choosing $\lambda_{\pi[i]}, i = d+1, \ldots, D$ to be contiguous in magnitude.

Exercise 2.11 (An EM Algorithm for PPCA) In Section 2.2.2 we showed that the ML estimate of the parameter $\theta = (\boldsymbol{\mu}, U_d, \sigma)$ of the PPCA model $\boldsymbol{x} = \boldsymbol{\mu} + U_d\boldsymbol{y} + \varepsilon$, where $\boldsymbol{y} \sim \mathcal{N}(0, I_d)$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_D)$, can be found in closed form, as shown in Theorem 2.8. An alternative approach, which can be advantageous for large D, is to view $\boldsymbol{y}$ as a hidden variable and use the EM algorithm described in Section B.2 to find the ML estimate. In this exercise, you will derive an EM algorithm for PPCA.

1. Show that the conditional distribution of the hidden variables given the observations is given by

$$\boldsymbol{y} \mid \boldsymbol{x} \sim \mathcal{N}(\Sigma_{\boldsymbol{x}}^{-1} U_d^\top (\boldsymbol{x} - \boldsymbol{\mu}), \sigma^2 \Sigma_{\boldsymbol{x}}^{-1}), \tag{2.100}$$

where $\Sigma_{\boldsymbol{x}} = U_d U_d^\top + \sigma^2 I_D$.

2. Let $w_i^k(\boldsymbol{y}) = p_{\theta^k}(\boldsymbol{y} \mid \boldsymbol{x}_i)$ be the posterior distribution of the hidden variables with parameters $\theta^k = (\boldsymbol{\mu}^k, U_d^k, \sigma^k)$ at iteration k of the EM algorithm. Show that the expected complete log-likelihood, $Q(\theta \mid \theta^k) = \mathbb{E}_{w^k}[\log p_\theta(\{\boldsymbol{x}_i\}_{i=1}^N, \{\boldsymbol{y}_j\}_{i=1}^N)]$, is given by:

$$\begin{aligned} -\sum_{i=1}^N \Big(\frac{D}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2} \big(\|\boldsymbol{x}_i - \boldsymbol{\mu}\|^2 - 2(\boldsymbol{x}_i - \boldsymbol{\mu})^\top U_d \langle \boldsymbol{y}_j \rangle^k \\ + \operatorname{trace} U_d^\top U_d \langle \boldsymbol{y}_j \boldsymbol{y}_i^\top \rangle^k \big) + \frac{1}{2} \operatorname{trace} \langle \boldsymbol{y}_i \boldsymbol{y}_i^\top \rangle^k \Big), \end{aligned} \tag{2.101}$$

where

$$\langle \boldsymbol{y}_i \rangle^k = \int_{\boldsymbol{y}} w_i^k(\boldsymbol{y}) \boldsymbol{y} \, d\boldsymbol{y} = \Sigma_{\boldsymbol{x}}^{k\,-1} U_d^{k\top} (\boldsymbol{x}_i - \boldsymbol{\mu}^k), \tag{2.102}$$

$$\langle \boldsymbol{y}_i \boldsymbol{y}_i^\top \rangle^k = \int_{\boldsymbol{y}} w_i^k(\boldsymbol{y}) \boldsymbol{y}\boldsymbol{y}^\top d\boldsymbol{y} = (\sigma^k)^2 \Sigma_{\boldsymbol{x}}^{k\,-1} + \langle \boldsymbol{y}_i \rangle^k \langle \boldsymbol{y}_i \rangle^{k\top}. \tag{2.103}$$

3. Show that the parameters $\theta = (\boldsymbol{\mu}, U_d, \sigma)$ that maximize $Q(\theta \mid \theta^k)$ are given by

$$\begin{bmatrix} U_d & \boldsymbol{\mu} \end{bmatrix} = \begin{bmatrix} \sum\limits_{i=1}^N \boldsymbol{x}_i \langle \boldsymbol{y}_i \rangle^{k\top} & \sum\limits_{i=1}^N \boldsymbol{x}_i \end{bmatrix} \begin{bmatrix} \sum\limits_{i=1}^N \langle \boldsymbol{y}_i \boldsymbol{y}_i^\top \rangle^k & \sum\limits_{i=1}^N \langle \boldsymbol{y}_i \rangle^k \\ \sum\limits_{i=1}^N \langle \boldsymbol{y}_i \rangle^{k\top} & N \end{bmatrix}^{-1}, \tag{2.104}$$

$$\sigma^2 = \frac{1}{ND} \sum_{i=1}^N \|\boldsymbol{x}_i - \boldsymbol{\mu}\|^2 - 2(\boldsymbol{x}_i - \boldsymbol{\mu})^\top U_d \langle \boldsymbol{y}_i \rangle^k + \operatorname{trace} U_d^\top U_d \langle \boldsymbol{y}_i \boldsymbol{y}_i^\top \rangle^k. \tag{2.105}$$

4. In practice, we know that the ML estimator for $\boldsymbol{\mu}$ is $\widehat{\boldsymbol{\mu}} = \frac{1}{N} \sum_{i=1}^N \boldsymbol{x}_i$. Therefore, a more efficient approach is to maximize $Q(\theta \mid \theta^k)$ only over the parameters (U_d, σ). Show that the optimal parameters are given by

$$U_d^{k+1} = \sum_{i=1}^N (\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}) \langle \boldsymbol{y}_i \rangle^{k\top} \Big(\sum_{i=1}^N \langle \boldsymbol{y}_i \boldsymbol{y}_i^\top \rangle^k \Big)^{-1}, \tag{2.106}$$

$$\sigma^{k+1} = \sqrt{\frac{1}{ND} \sum_{i=1}^N \|\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}\|^2 - 2(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})^\top U_d^{k+1} \langle \boldsymbol{y}_i \rangle^k + \operatorname{trace} U_d^{k+1\top} U_d^{k+1} \langle \boldsymbol{y}_i \boldsymbol{y}_i^\top \rangle^k}.$$

where $\langle \boldsymbol{y}_i \rangle^k$ is computed with $\boldsymbol{\mu}^k = \widehat{\boldsymbol{\mu}}$. Show also that the above iterations can be re-written as

$$U_d^{k+1} = \widehat{\Sigma}_N U_d^k \big[(\sigma^k)^2 I_d + \Sigma_{\boldsymbol{x}}^{k\,-1} U_d^{k\top} \widehat{\Sigma}_N U_d^k \big]^{-1}, \tag{2.107}$$

$$\sigma^{k+1} = \sqrt{\frac{1}{D} \operatorname{trace}(\widehat{\Sigma}_N - \widehat{\Sigma}_N U_d^k \Sigma_{\boldsymbol{x}}^{k\,-1} U_d^{k+1\top})}, \tag{2.108}$$

where $\widehat{\Sigma}_N = \frac{1}{N} \sum_{i=1}^N (\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})^\top$.