Chapter 23

Applications of SVD and Pseudo-Inverses

> De tous les principes qu'on peut proposer pour cet objet, je pense qu'il n'en est pas de plus général, de plus exact, ni d'une application plus facile, que celui dont nous avons fait usage dans les recherches précédentes, et qui consiste à rendre *minimum* la somme des carrés des erreurs. Par ce moyen il s'établit entre les erreurs une sorte d'équilibre qui, empêchant les extrêmes de prévaloir, est très propre às faire connaitre l'état du système le plus proche de la vérité.
>
> —**Legendre, 1805,** *Nouvelles Méthodes pour la détermination des Orbites des Comètes*

23.1 Least Squares Problems and the Pseudo-Inverse

This chapter presents several applications of SVD. The first one is the pseudo-inverse, which plays a crucial role in solving linear systems by the method of least squares. The second application is data compression. The third application is principal component analysis (PCA), whose purpose is to identify patterns in data and understand the variance–covariance structure of the data. The fourth application is the best affine approximation of a set of data, a problem closely related to PCA.

The method of least squares is a way of "solving" an overdetermined system of linear equations

$$Ax = b,$$

i.e., a system in which A is a rectangular $m \times n$ matrix with more equations than unknowns (when $m > n$). Historically, the method of least squares was used by Gauss and Legendre to solve problems in astronomy and geodesy. The method was first published by Legendre in 1805 in a paper on methods for determining the orbits of comets. However, Gauss had already used the method of least squares as early as 1801 to determine the orbit of the asteroid

Ceres, and he published a paper about it in 1810 after the discovery of the asteroid Pallas. Incidentally, it is in that same paper that Gaussian elimination using pivots is introduced.

The reason why more equations than unknowns arise in such problems is that repeated measurements are taken to minimize errors. This produces an overdetermined and often inconsistent system of linear equations. For example, Gauss solved a system of eleven equations in six unknowns to determine the orbit of the asteroid Pallas.

Example 23.1. As a concrete illustration, suppose that we observe the motion of a small object, assimilated to a point, in the plane. From our observations, we suspect that this point moves along a straight line, say of equation $y = cx + d$. Suppose that we observed the moving point at three different locations (x_1, y_1), (x_2, y_2), and (x_3, y_3). Then we should have

$$\begin{aligned} d + cx_1 &= y_1, \\ d + cx_2 &= y_2, \\ d + cx_3 &= y_3. \end{aligned}$$

If there were no errors in our measurements, these equations would be compatible, and c and d would be determined by only two of the equations. However, in the presence of errors, the system may be inconsistent. Yet we would like to find c and d!

The idea of the method of least squares is to determine (c, d) such that it minimizes the sum of the squares of the errors, namely,

$$(d + cx_1 - y_1)^2 + (d + cx_2 - y_2)^2 + (d + cx_3 - y_3)^2.$$

See Figure 23.1.

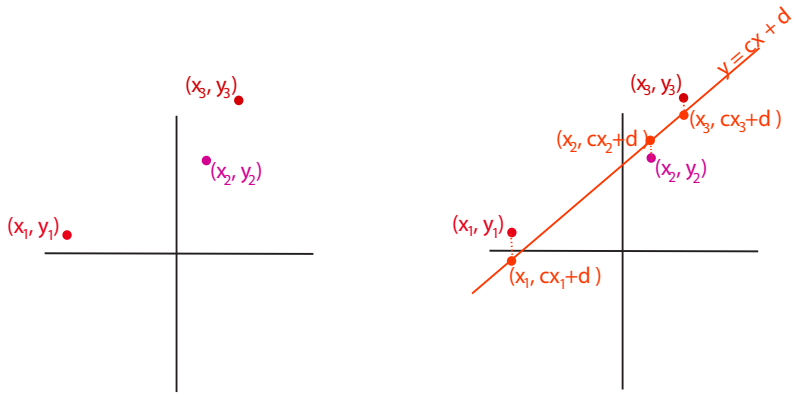

Figure 23.1: Given three points (x_1, y_1), (x_2, y_2), (x_3, y_3), we want to determine the line $y = cx + d$ which minimizes the lengths of the dashed vertical lines.

In general, for an overdetermined $m \times n$ system $Ax = b$, what Gauss and Legendre discovered is that there are solutions x minimizing

$$\|Ax - b\|_2^2$$

(where $\|u\|_2^2 = u_1^2 + \cdots + u_n^2$, the square of the Euclidean norm of the vector $u = (u_1, \ldots, u_n)$), and that these solutions are given by the square $n \times n$ system

$$A^\top Ax = A^\top b,$$

called the *normal equations*. Furthermore, when the columns of A are linearly independent, it turns out that $A^\top A$ is invertible, and so x is unique and given by

$$x = (A^\top A)^{-1} A^\top b.$$

Note that $A^\top A$ is a symmetric matrix, one of the nice features of the normal equations of a least squares problem. For instance, since the above problem in matrix form is represented as

$$\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \end{pmatrix} \begin{pmatrix} d \\ c \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix},$$

the normal equations are

$$\begin{pmatrix} 3 & x_1 + x_2 + x_3 \\ x_1 + x_2 + x_3 & x_1^2 + x_2^2 + x_3^2 \end{pmatrix} \begin{pmatrix} d \\ c \end{pmatrix} = \begin{pmatrix} y_1 + y_2 + y_3 \\ x_1 y_1 + x_2 y_2 + x_3 y_3 \end{pmatrix}.$$

In fact, given any real $m \times n$ matrix A, there is always a unique x^+ of minimum norm that minimizes $\|Ax - b\|_2^2$, even when the columns of A are linearly dependent. How do we prove this, and how do we find x^+?

Theorem 23.1. *Every linear system $Ax = b$, where A is an $m \times n$ matrix, has a unique least squares solution x^+ of smallest norm.*

Proof. Geometry offers a nice proof of the existence and uniqueness of x^+. Indeed, we can interpret b as a point in the Euclidean (affine) space $\mathbb{R}^m$, and the image subspace of A (also called the column space of A) as a subspace U of $\mathbb{R}^m$ (passing through the origin). Then it is clear that

$$\inf_{x \in \mathbb{R}^n} \|Ax - b\|_2^2 = \inf_{y \in U} \|y - b\|_2^2,$$

with $U = \operatorname{Im} A$, and we claim that x minimizes $\|Ax - b\|_2^2$ iff $Ax = p$, where p the orthogonal projection of b onto the subspace U.

Recall from Section 13.1 that the orthogonal projection $p_U \colon U \oplus U^\perp \to U$ is the linear map given by

$$p_U(u + v) = u,$$

with $u \in U$ and $v \in U^\perp$. If we let $p = p_U(b) \in U$, then for any point $y \in U$, the vectors $\overrightarrow{py} = y - p \in U$ and $\overrightarrow{bp} = p - b \in U^\perp$ are orthogonal, which implies that

$$\|\overrightarrow{by}\|_2^2 = \|\overrightarrow{bp}\|_2^2 + \|\overrightarrow{py}\|_2^2,$$

where $\overrightarrow{by} = y - b$. Thus, p is indeed the unique point in U that minimizes the distance from b to any point in U. See Figure 23.2.

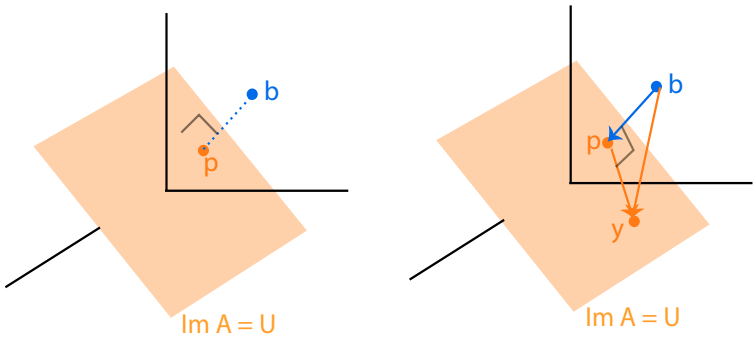

Figure 23.2: Given a 3×2 matrix A, $U = \operatorname{Im} A$ is the peach plane in $\mathbb{R}^3$ and p is the orthogonal projection of b onto U. Furthermore, given $y \in U$, the points b, y, and p are the vertices of a right triangle.

Thus the problem has been reduced to proving that there is a unique x^+ of minimum norm such that $Ax^+ = p$, with $p = p_U(b) \in U$, the orthogonal projection of b onto U. We use the fact that

$$\mathbb{R}^n = \operatorname{Ker} A \oplus (\operatorname{Ker} A)^\perp.$$

Consequently, every $x \in \mathbb{R}^n$ can be written uniquely as $x = u + v$, where $u \in \operatorname{Ker} A$ and $v \in (\operatorname{Ker} A)^\perp$, and since u and v are orthogonal,

$$\|x\|_2^2 = \|u\|_2^2 + \|v\|_2^2.$$

Furthermore, since $u \in \operatorname{Ker} A$, we have $Au = 0$, and thus $Ax = p$ iff $Av = p$, which shows that the solutions of $Ax = p$ for which x has minimum norm must belong to $(\operatorname{Ker} A)^\perp$. However, the restriction of A to $(\operatorname{Ker} A)^\perp$ is injective. This is because if $Av_1 = Av_2$, where $v_1, v_2 \in (\operatorname{Ker} A)^\perp$, then $A(v_2 - v_1) = 0$, which implies $v_2 - v_1 \in \operatorname{Ker} A$, and since $v_1, v_2 \in (\operatorname{Ker} A)^\perp$, we also have $v_2 - v_1 \in (\operatorname{Ker} A)^\perp$, and consequently, $v_2 - v_1 = 0$. This shows that there is a unique x^+ of minimum norm such that $Ax^+ = p$, and that x^+ must belong to $(\operatorname{Ker} A)^\perp$. By our previous reasoning, x^+ is the unique vector of minimum norm minimizing $\|Ax - b\|_2^2$. □

The proof also shows that x minimizes $\|Ax - b\|_2^2$ iff $\overrightarrow{pb} = b - Ax$ is orthogonal to U, which can be expressed by saying that $b - Ax$ is orthogonal to every column of A. However, this is equivalent to

$$A^\top(b - Ax) = 0, \quad \text{i.e.,} \quad A^\top Ax = A^\top b.$$

Finally, it turns out that the minimum norm least squares solution x^+ can be found in terms of the pseudo-inverse A^+ of A, which is itself obtained from any SVD of A.

Definition 23.1. Given any nonzero $m \times n$ matrix A of rank r, if $A = VDU^\top$ is an SVD of A such that

$$D = \begin{pmatrix} \Lambda & 0_{r,n-r} \\ 0_{m-r,r} & 0_{m-r,n-r} \end{pmatrix},$$

with

$$\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_r)$$

an $r \times r$ diagonal matrix consisting of the nonzero singular values of A, then if we let D^+ be the $n \times m$ matrix

$$D^+ = \begin{pmatrix} \Lambda^{-1} & 0_{r,m-r} \\ 0_{n-r,r} & 0_{n-r,m-r} \end{pmatrix},$$

with

$$\Lambda^{-1} = \mathrm{diag}(1/\lambda_1, \ldots, 1/\lambda_r),$$

the *pseudo-inverse* of A is defined by

$$A^+ = UD^+V^\top.$$

If $A = 0_{m,n}$ is the zero matrix, we set $A^+ = 0_{n,m}$. Observe that D^+ is obtained from D by inverting the nonzero diagonal entries of D, leaving all zeros in place, and then transposing the matrix. For example, given the matrix

$$D = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

its pseudo-inverse is

$$D^+ = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

The pseudo-inverse of a matrix is also known as the *Moore–Penrose pseudo-inverse*.

Actually, it seems that A^+ depends on the specific choice of U and V in an SVD (U, D, V) for A, but the next theorem shows that this is not so.

Theorem 23.2. *The least squares solution of smallest norm of the linear system $Ax = b$, where A is an $m \times n$ matrix, is given by*

$$x^+ = A^+ b = UD^+V^\top b.$$

Proof. First assume that A is a (rectangular) diagonal matrix D, as above. Then since x minimizes $\|Dx - b\|_2^2$ iff Dx is the projection of b onto the image subspace F of D, it is fairly obvious that $x^+ = D^+ b$. Otherwise, we can write

$$A = VDU^\top,$$

where U and V are orthogonal. However, since V is an isometry,

$$\|Ax - b\|_2 = \|VDU^\top x - b\|_2 = \|DU^\top x - V^\top b\|_2.$$

Letting $y = U^\top x$, we have $\|x\|_2 = \|y\|_2$, since U is an isometry, and since U is surjective, $\|Ax - b\|_2$ is minimized iff $\|Dy - V^\top b\|_2$ is minimized, and we have shown that the least solution is

$$y^+ = D^+V^\top b.$$

Since $y = U^\top x$, with $\|x\|_2 = \|y\|_2$, we get

$$x^+ = UD^+V^\top b = A^+ b.$$

Thus, the pseudo-inverse provides the optimal solution to the least squares problem. □

By Theorem 23.2 and Theorem 23.1, $A^+ b$ is uniquely defined by every b, and thus A^+ depends only on A.

The `Matlab` command for computing the pseudo-inverse B of the matrix A is
`B = pinv(A)`.

Example 23.2. If A is the rank 2 matrix

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 5 \\ 3 & 4 & 5 & 6 \\ 4 & 5 & 6 & 7 \end{pmatrix}$$

whose eigenvalues are $-1.1652, 0, 0, 17.1652$, using `Matlab` we obtain the SVD $A = VDU^\top$ with

$$U = \begin{pmatrix} -0.3147 & 0.7752 & 0.2630 & -0.4805 \\ -0.4275 & 0.3424 & 0.0075 & 0.8366 \\ -0.5402 & -0.0903 & -0.8039 & -0.2319 \\ -0.6530 & -0.5231 & 0.5334 & -0.1243 \end{pmatrix},$$

$$V = \begin{pmatrix} -0.3147 & -0.7752 & 0.5452 & 0.0520 \\ -0.4275 & -0.3424 & -0.7658 & 0.3371 \\ -0.5402 & 0.0903 & -0.1042 & -0.8301 \\ -0.6530 & 0.5231 & 0.3247 & 0.4411 \end{pmatrix}, \quad D = \begin{pmatrix} 17.1652 & 0 & 0 & 0 \\ 0 & 1.1652 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Then

$$D^+ = \begin{pmatrix} 0.0583 & 0 & 0 & 0 \\ 0 & 0.8583 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

and

$$A^+ = UD^+V^\top = \begin{pmatrix} -0.5100 & -0.2200 & 0.0700 & 0.3600 \\ -0.2200 & -0.0900 & 0.0400 & 0.1700 \\ 0.0700 & 0.0400 & 0.0100 & -0.0200 \\ 0.3600 & 0.1700 & -0.0200 & -0.2100 \end{pmatrix},$$

which is also the result obtained by calling `pinv(A)`.

If A is an $m \times n$ matrix of rank n (and so $m \geq n$), it is immediately shown that the QR-decomposition in terms of Householder transformations applies as follows:

There are n $m \times m$ matrices $H_1, \ldots, H_n$, Householder matrices or the identity, and an upper triangular $m \times n$ matrix R of rank n such that

$$A = H_1 \cdots H_n R.$$

Then because each H_i is an isometry,

$$\|Ax - b\|_2 = \|Rx - H_n \cdots H_1 b\|_2,$$

and the least squares problem $Ax = b$ is equivalent to the system

$$Rx = H_n \cdots H_1 b.$$

Now the system

$$Rx = H_n \cdots H_1 b$$

is of the form

$$\begin{pmatrix} R_1 \\ 0_{m-n} \end{pmatrix} x = \begin{pmatrix} c \\ d \end{pmatrix},$$

where R_1 is an invertible $n \times n$ matrix (since A has rank n), $c \in \mathbb{R}^n$, and $d \in \mathbb{R}^{m-n}$, and the least squares solution of smallest norm is

$$x^+ = R_1^{-1} c.$$

Since R_1 is a triangular matrix, it is very easy to invert R_1.

The method of least squares is one of the most effective tools of the mathematical sciences. There are entire books devoted to it. Readers are advised to consult Strang [170], Golub and Van Loan [80], Demmel [48], and Trefethen and Bau [176], where extensions and applications of least squares (such as weighted least squares and recursive least squares) are described. Golub and Van Loan [80] also contains a very extensive bibliography, including a list of books on least squares.

23.2 Properties of the Pseudo-Inverse

We begin this section with a proposition which provides a way to calculate the pseudo-inverse of an $m \times n$ matrix A without first determining an SVD factorization.

Proposition 23.3. *When A has full rank, the pseudo-inverse A^+ can be expressed as $A^+ = (A^\top A)^{-1}A^\top$ when $m \geq n$, and as $A^+ = A^\top(AA^\top)^{-1}$ when $n \geq m$. In the first case ($m \geq n$), observe that $A^+A = I$, so A^+ is a left inverse of A; in the second case ($n \geq m$), we have $AA^+ = I$, so A^+ is a right inverse of A.*

Proof. If $m \geq n$ and A has full rank n, we have

$$A = V \begin{pmatrix} \Lambda \\ 0_{m-n,n} \end{pmatrix} U^\top$$

with Λ an $n \times n$ diagonal invertible matrix (with positive entries), so

$$A^+ = U \begin{pmatrix} \Lambda^{-1} & 0_{n,m-n} \end{pmatrix} V^\top.$$

We find that

$$A^\top A = U \begin{pmatrix} \Lambda & 0_{n,m-n} \end{pmatrix} V^\top V \begin{pmatrix} \Lambda \\ 0_{m-n,n} \end{pmatrix} U^\top = U\Lambda^2 U^\top,$$

which yields

$$(A^\top A)^{-1}A^\top = U\Lambda^{-2}U^\top U \begin{pmatrix} \Lambda & 0_{n,m-n} \end{pmatrix} V^\top = U \begin{pmatrix} \Lambda^{-1} & 0_{n,m-n} \end{pmatrix} V^\top = A^+.$$

Therefore, if $m \geq n$ and A has full rank n, then

$$A^+ = (A^\top A)^{-1}A^\top.$$

If $n \geq m$ and A has full rank m, then

$$A = V \begin{pmatrix} \Lambda & 0_{m,n-m} \end{pmatrix} U^\top$$

with Λ an $m \times m$ diagonal invertible matrix (with positive entries), so

$$A^+ = U \begin{pmatrix} \Lambda^{-1} \\ 0_{n-m,m} \end{pmatrix} V^\top.$$

We find that

$$AA^\top = V \begin{pmatrix} \Lambda & 0_{m,n-m} \end{pmatrix} U^\top U \begin{pmatrix} \Lambda \\ 0_{n-m,m} \end{pmatrix} V^\top = V\Lambda^2 V^\top,$$

which yields

$$A^\top(AA^\top)^{-1} = U \begin{pmatrix} \Lambda \\ 0_{n-m,m} \end{pmatrix} V^\top V\Lambda^{-2}V^\top = U \begin{pmatrix} \Lambda^{-1} \\ 0_{n-m,m} \end{pmatrix} V^\top = A^+.$$

Therefore, if $n \geq m$ and A has full rank m, then $A^+ = A^\top(AA^\top)^{-1}$. □

For example, if $A = \begin{pmatrix} 1 & 2 \\ 2 & 3 \\ 0 & 1 \end{pmatrix}$, then A has rank 2 and since $m \geq n$, $A^+ = (A^\top A)^{-1} A^\top$ where

$$A^+ = \begin{pmatrix} 5 & 8 \\ 8 & 14 \end{pmatrix}^{-1} A^\top = \begin{pmatrix} 7/3 & -4/3 \\ 4/3 & 5/6 \end{pmatrix} \begin{pmatrix} 1 & 2 & 0 \\ 2 & 3 & 1 \end{pmatrix} = \begin{pmatrix} -1/3 & 2/3 & -4/3 \\ 1/3 & -1/6 & 5/6 \end{pmatrix}.$$

If $A = \begin{pmatrix} 1 & 2 & 3 & 0 \\ 0 & 1 & 1 & -1 \end{pmatrix}$, since A has rank 2 and $n \geq m$, then $A^+ = A^\top (AA^\top)^{-1}$ where

$$A^+ = A^\top \begin{pmatrix} 14 & 5 \\ 5 & 3 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & 0 \\ 2 & 1 \\ 3 & 1 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} 3/17 & -5/17 \\ -5/17 & 14/17 \end{pmatrix} = \begin{pmatrix} 3/17 & -5/17 \\ 1/17 & 4/17 \\ 4/17 & -1/17 \\ 5/17 & -14/17 \end{pmatrix}.$$

Let $A = V\Sigma U^\top$ be an SVD for any $m \times n$ matrix A. It is easy to check that both AA^+ and A^+A are symmetric matrices. In fact,

$$AA^+ = V\Sigma U^\top U \Sigma^+ V^\top = V \Sigma \Sigma^+ V^\top = V \begin{pmatrix} I_r & 0 \\ 0 & 0_{m-r} \end{pmatrix} V^\top$$

and

$$A^+A = U\Sigma^+ V^\top V \Sigma U^\top = U \Sigma^+ \Sigma U^\top = U \begin{pmatrix} I_r & 0 \\ 0 & 0_{n-r} \end{pmatrix} U^\top.$$

From the above expressions we immediately deduce that

$$\begin{aligned} AA^+A &= A, \\ A^+AA^+ &= A^+, \end{aligned}$$

and that

$$\begin{aligned} (AA^+)^2 &= AA^+, \\ (A^+A)^2 &= A^+A, \end{aligned}$$

so both AA^+ and A^+A are orthogonal projections (since they are both symmetric).

Proposition 23.4. *The matrix AA^+ is the orthogonal projection onto the range of A and A^+A is the orthogonal projection onto* $\mathrm{Ker}(A)^\perp = \mathrm{Im}(A^\top)$*, the range of $A^\top$.*

Proof. Obviously, we have $\mathrm{range}(AA^+) \subseteq \mathrm{range}(A)$, and for any $y = Ax \in \mathrm{range}(A)$, since $AA^+A = A$, we have

$$AA^+y = AA^+Ax = Ax = y,$$

so the image of AA^+ is indeed the range of A. It is also clear that $\mathrm{Ker}(A) \subseteq \mathrm{Ker}(A^+A)$, and since $AA^+A = A$, we also have $\mathrm{Ker}(A^+A) \subseteq \mathrm{Ker}(A)$, and so

$$\mathrm{Ker}(A^+A) = \mathrm{Ker}(A).$$

Since A^+A is symmetric, $\mathrm{range}(A^+A) = \mathrm{range}((A^+A)^\top) = \mathrm{Ker}(A^+A)^\perp = \mathrm{Ker}(A)^\perp$, as claimed. □

Proposition 23.5. *The set* $\mathrm{range}(A) = \mathrm{range}(AA^+)$ *consists of all vectors* $y \in \mathbb{R}^m$ *such that*

$$V^\top y = \begin{pmatrix} z \\ 0 \end{pmatrix},$$

with $z \in \mathbb{R}^r$.

Proof. Indeed, if $y = Ax$, then

$$V^\top y = V^\top Ax = V^\top V \Sigma U^\top x = \Sigma U^\top x = \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0_{m-r} \end{pmatrix} U^\top x = \begin{pmatrix} z \\ 0 \end{pmatrix},$$

where Σ_r is the $r \times r$ diagonal matrix $\mathrm{diag}(\sigma_1, \ldots, \sigma_r)$. Conversely, if $V^\top y = \left(\begin{smallmatrix} z \\ 0 \end{smallmatrix}\right)$, then $y = V \left(\begin{smallmatrix} z \\ 0 \end{smallmatrix}\right)$, and

$$\begin{aligned} AA^+ y &= V \begin{pmatrix} I_r & 0 \\ 0 & 0_{m-r} \end{pmatrix} V^\top y \\ &= V \begin{pmatrix} I_r & 0 \\ 0 & 0_{m-r} \end{pmatrix} V^\top V \begin{pmatrix} z \\ 0 \end{pmatrix} \\ &= V \begin{pmatrix} I_r & 0 \\ 0 & 0_{m-r} \end{pmatrix} \begin{pmatrix} z \\ 0 \end{pmatrix} \\ &= V \begin{pmatrix} z \\ 0 \end{pmatrix} = y, \end{aligned}$$

which shows that y belongs to the range of A. □

Similarly, we have the following result.

Proposition 23.6. *The set* $\mathrm{range}(A^+A) = \mathrm{Ker}(A)^\perp$ *consists of all vectors* $y \in \mathbb{R}^n$ *such that*

$$U^\top y = \begin{pmatrix} z \\ 0 \end{pmatrix},$$

with $z \in \mathbb{R}^r$.

Proof. If $y = A^+Au$, then

$$y = A^+Au = U\begin{pmatrix} I_r & 0 \\ 0 & 0_{n-r} \end{pmatrix} U^\top u = U\begin{pmatrix} z \\ 0 \end{pmatrix},$$

for some $z \in \mathbb{R}^r$. Conversely, if $U^\top y = \binom{z}{0}$, then $y = U\binom{z}{0}$, and so

$$\begin{aligned} A^+AU\begin{pmatrix} z \\ 0 \end{pmatrix} &= U\begin{pmatrix} I_r & 0 \\ 0 & 0_{n-r} \end{pmatrix} U^\top U\begin{pmatrix} z \\ 0 \end{pmatrix} \\ &= U\begin{pmatrix} I_r & 0 \\ 0 & 0_{n-r} \end{pmatrix}\begin{pmatrix} z \\ 0 \end{pmatrix} \\ &= U\begin{pmatrix} z \\ 0 \end{pmatrix} = y, \end{aligned}$$

which shows that $y \in \text{range}(A^+A)$. □

Analogous results hold for complex matrices, but in this case, V and U are unitary matrices and AA^+ and A^+A are Hermitian orthogonal projections.

If A is a normal matrix, which means that $AA^\top = A^\top A$, then there is an intimate relationship between SVD's of A and block diagonalizations of A. As a consequence, the pseudo-inverse of a normal matrix A can be obtained directly from a block diagonalization of A.

If A is a (real) normal matrix, then we know from Theorem 17.18 that A can be block diagonalized with respect to an orthogonal matrix U as

$$A = U\Lambda U^\top,$$

where Λ is the (real) block diagonal matrix

$$\Lambda = \text{diag}(B_1, \ldots, B_n),$$

consisting either of 2×2 blocks of the form

$$B_j = \begin{pmatrix} \lambda_j & -\mu_j \\ \mu_j & \lambda_j \end{pmatrix}$$

with $\mu_j \neq 0$, or of one-dimensional blocks $B_k = (\lambda_k)$. Then we have the following proposition:

Proposition 23.7. *For any (real) normal matrix A and any block diagonalization $A = U\Lambda U^\top$ of A as above, the pseudo-inverse of A is given by*

$$A^+ = U\Lambda^+U^\top,$$

where Λ^+ is the pseudo-inverse of Λ. Furthermore, if

$$\Lambda = \begin{pmatrix} \Lambda_r & 0 \\ 0 & 0 \end{pmatrix},$$

where Λ_r has rank r, then

$$\Lambda^+ = \begin{pmatrix} \Lambda_r^{-1} & 0 \\ 0 & 0 \end{pmatrix}.$$

Proof. Assume that $B_1, \ldots, B_p$ are 2×2 blocks and that $\lambda_{2p+1}, \ldots, \lambda_n$ are the scalar entries. We know that the numbers $\lambda_j \pm i\mu_j$, and the λ_{2p+k} are the eigenvalues of A. Let $\rho_{2j-1} = \rho_{2j} = \sqrt{\lambda_j^2 + \mu_j^2} = \sqrt{\det(B_i)}$ for $j = 1, \ldots, p$, $\rho_j = |\lambda_j|$ for $j = 2p+1, \ldots, r$. Multiplying U by a suitable permutation matrix, we may assume that the blocks of Λ are ordered so that $\rho_1 \geq \rho_2 \geq \cdots \geq \rho_r > 0$. Then it is easy to see that

$$AA^\top = A^\top A = U\Lambda U^\top U\Lambda^\top U^\top = U\Lambda\Lambda^\top U^\top,$$

with

$$\Lambda\Lambda^\top = \mathrm{diag}(\rho_1^2, \ldots, \rho_r^2, 0, \ldots, 0),$$

so $\rho_1 \geq \rho_2 \geq \cdots \geq \rho_r > 0$ are the singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ of A. Define the diagonal matrix

$$\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_r, 0, \ldots, 0),$$

where $r = \mathrm{rank}(A)$, $\sigma_1 \geq \cdots \geq \sigma_r > 0$ and the block diagonal matrix Θ defined such that the block B_i in Λ is replaced by the block $\sigma^{-1}B_i$ where $\sigma = \sqrt{\det(B_i)}$, the nonzero scalar λ_j is replaced $\lambda_j/|\lambda_j|$, and a diagonal zero is replaced by 1. Observe that Θ is an orthogonal matrix and

$$\Lambda = \Theta\Sigma.$$

But then we can write

$$A = U\Lambda U^\top = U\Theta\Sigma U^\top,$$

and we if let $V = U\Theta$, since U is orthogonal and Θ is also orthogonal, V is also orthogonal and $A = V\Sigma U^\top$ *is an SVD for* A. Now we get

$$A^+ = U\Sigma^+ V^\top = U\Sigma^+\Theta^\top U^\top.$$

However, since Θ is an orthogonal matrix, $\Theta^\top = \Theta^{-1}$, and a simple calculation shows that

$$\Sigma^+\Theta^\top = \Sigma^+\Theta^{-1} = \Lambda^+,$$

which yields the formula

$$A^+ = U\Lambda^+ U^\top.$$

Also observe that Λ_r is invertible and

$$\Lambda^+ = \begin{pmatrix} \Lambda_r^{-1} & 0 \\ 0 & 0 \end{pmatrix}.$$

Therefore, the pseudo-inverse of a normal matrix can be computed directly from any block diagonalization of A, as claimed. □

Example 23.3. Consider the following real diagonal form of the normal matrix

$$A = \begin{pmatrix} -2.7500 & 2.1651 & -0.8660 & 0.5000 \\ 2.1651 & -0.2500 & -1.5000 & 0.8660 \\ 0.8660 & 1.5000 & 0.7500 & -0.4330 \\ -0.5000 & -0.8660 & -0.4330 & 0.2500 \end{pmatrix} = U\Lambda U^\top,$$

with

$$U = \begin{pmatrix} \cos(\pi/3) & 0 & \sin(\pi/3) & 0 \\ \sin(\pi/3) & 0 & -\cos(\pi/3) & 0 \\ 0 & \cos(\pi/6) & 0 & \sin(\pi/6) \\ 0 & -\cos(\pi/6) & 0 & \sin(\pi/6) \end{pmatrix}, \quad \Lambda = \begin{pmatrix} 1 & -2 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 0 & 0 & -4 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

We obtain

$$\Lambda^+ = \begin{pmatrix} 1/5 & 2/5 & 0 & 0 \\ -2/5 & 1/5 & 0 & 0 \\ 0 & 0 & -1/4 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

and the pseudo-inverse of A is

$$A^+ = U\Lambda^+ U^\top = \begin{pmatrix} -0.1375 & 0.1949 & 0.1732 & -0.1000 \\ 0.1949 & 0.0875 & 0.3000 & -0.1732 \\ -0.1732 & -0.3000 & 0.1500 & -0.0866 \\ 0.1000 & 0.1732 & -0.0866 & 0.0500 \end{pmatrix},$$

which agrees with `pinv(A)`.

The following properties, due to Penrose, characterize the pseudo-inverse of a matrix. We have already proved that the pseudo-inverse satisfies these equations. For a proof of the converse, see Kincaid and Cheney [102].

Proposition 23.8. *Given any $m \times n$ matrix A (real or complex), the pseudo-inverse A^+ of A is the unique $n \times m$ matrix satisfying the following properties:*

$$\begin{aligned} AA^+A &= A, \\ A^+AA^+ &= A^+, \\ (AA^+)^\top &= AA^+, \\ (A^+A)^\top &= A^+A. \end{aligned}$$

23.3 Data Compression and SVD

Among the many applications of SVD, a very useful one is *data compression*, notably for images. In order to make precise the notion of closeness of matrices, we use the notion of

matrix norm. This concept is defined in Chapter 9, and the reader may want to review it before reading any further.

Given an $m \times n$ matrix of rank r, we would like to find a best approximation of A by a matrix B of rank $k \leq r$ (actually, $k < r$) such that $\|A - B\|_2$ (or $\|A - B\|_F$) is minimized. The following proposition is known as the *Eckart–Young theorem*.

Proposition 23.9. *Let A be an $m \times n$ matrix of rank r and let $VDU^\top = A$ be an SVD for A. Write u_i for the columns of U, v_i for the columns of V, and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p$ for the singular values of A ($p = \min(m, n)$). Then a matrix of rank $k < r$ closest to A (in the $\| \|_2$ norm) is given by*

$$A_k = \sum_{i=1}^{k} \sigma_i v_i u_i^\top = V \mathrm{diag}(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0) U^\top$$

and $\|A - A_k\|_2 = \sigma_{k+1}$.

Proof. By construction, A_k has rank k, and we have

$$\|A - A_k\|_2 = \Big\| \sum_{i=k+1}^{p} \sigma_i v_i u_i^\top \Big\|_2 = \big\| V \mathrm{diag}(0, \ldots, 0, \sigma_{k+1}, \ldots, \sigma_p) U^\top \big\|_2 = \sigma_{k+1}.$$

It remains to show that $\|A - B\|_2 \geq \sigma_{k+1}$ for all rank k matrices B. Let B be any rank k matrix, so its kernel has dimension $n - k$. The subspace U_{k+1} spanned by $(u_1, \ldots, u_{k+1})$ has dimension $k + 1$, and because the sum of the dimensions of the kernel of B and of U_{k+1} is $(n - k) + k + 1 = n + 1$, these two subspaces must intersect in a subspace of dimension at least 1. Pick any unit vector h in $\mathrm{Ker}(B) \cap U_{k+1}$. Then since $Bh = 0$, and since U and V are isometries, we have

$$\|A - B\|_2^2 \geq \|(A - B)h\|_2^2 = \|Ah\|_2^2 = \big\|VDU^\top h\big\|_2^2 = \big\|DU^\top h\big\|_2^2 \geq \sigma_{k+1}^2 \big\|U^\top h\big\|_2^2 = \sigma_{k+1}^2,$$

which proves our claim. □

Note that A_k can be stored using $(m + n)k$ entries, as opposed to mn entries. When $k \ll m$, this is a substantial gain.

Example 23.4. Consider the badly conditioned symmetric matrix

$$A = \begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix}$$

from Section 9.5. Since A is SPD, we have the SVD

$$A = UDU^\top,$$

with

$$U = \begin{pmatrix} -0.5286 & -0.6149 & 0.3017 & -0.5016 \\ -0.3803 & -0.3963 & -0.0933 & 0.8304 \\ -0.5520 & 0.2716 & -0.7603 & -0.2086 \\ -0.5209 & 0.6254 & 0.5676 & 0.1237 \end{pmatrix}, D = \begin{pmatrix} 30.2887 & 0 & 0 & 0 \\ 0 & 3.8581 & 0 & 0 \\ 0 & 0 & 0.8431 & 0 \\ 0 & 0 & 0 & 0.0102 \end{pmatrix}.$$

If we set $\sigma_3 = \sigma_4 = 0$, we obtain the best rank 2 approximation

$$A_2 = U(:, 1:2) * D(:, 1:2) * U(:, 1:2)' = \begin{pmatrix} 9.9207 & 7.0280 & 8.1923 & 6.8563 \\ 7.0280 & 4.9857 & 5.9419 & 5.0436 \\ 8.1923 & 5.9419 & 9.5122 & 9.3641 \\ 6.8563 & 5.0436 & 9.3641 & 9.7282 \end{pmatrix}.$$

A nice example of the use of Proposition 23.9 in image compression is given in Demmel [48], Chapter 3, Section 3.2.3, pages 113–115; see the Matlab demo.

Proposition 23.9 also holds for the Frobenius norm; see Problem 23.4.

An interesting topic that we have not addressed is the actual computation of an SVD. This is a very interesting but tricky subject. Most methods reduce the computation of an SVD to the diagonalization of a well-chosen symmetric matrix which is not $A^\top A$; see Problem 22.1 and Problem 22.3. Interested readers should read Section 5.4 of Demmel's excellent book [48], which contains an overview of most known methods and an extensive list of references.

23.4 Principal Components Analysis (PCA)

Suppose we have a set of data consisting of n points $X_1, \ldots, X_n$, with each $X_i \in \mathbb{R}^d$ *viewed as a row vector*. Think of the X_i's as persons, and if $X_i = (x_{i1}, \ldots, x_{id})$, each x_{ij} is the value of some *feature* (or *attribute*) of that person.

Example 23.5. For example, the X_i's could be mathematicians, $d = 2$, and the first component, x_{i1}, of X_i could be the year that X_i was born, and the second component, x_{i2}, the length of the beard of X_i in centimeters. Here is a small data set.

Name	year	length
Carl Friedrich Gauss	1777	0
Camille Jordan	1838	12
Adrien-Marie Legendre	1752	0
Bernhard Riemann	1826	15
David Hilbert	1862	2
Henri Poincaré	1854	5
Emmy Noether	1882	0
Karl Weierstrass	1815	0
Eugenio Beltrami	1835	2
Hermann Schwarz	1843	20

We usually form the $n \times d$ matrix X whose ith row is X_i, with $1 \leq i \leq n$. Then the jth column is denoted by C_j $(1 \leq j \leq d)$. It is sometimes called a *feature vector*, but this terminology is far from being universally accepted. In fact, many people in computer vision call the data points X_i feature vectors!

The purpose of *principal components analysis*, for short *PCA*, is to identify patterns in data and understand the *variance–covariance* structure of the data. This is useful for the following tasks:

1. Data reduction: Often much of the variability of the data can be accounted for by a smaller number of *principal components*.

2. Interpretation: PCA can show relationships that were not previously suspected.

Given a vector (a *sample* of measurements) $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, recall that the *mean* (or *average*) $\overline{x}$ of x is given by

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

We let $x - \overline{x}$ denote the *centered data point*

$$x - \overline{x} = (x_1 - \overline{x}, \ldots, x_n - \overline{x}).$$

In order to *measure the spread* of the x_i's around the mean, we define the *sample variance* (for short, *variance*) $\mathrm{var}(x)$ (or s^2) of the sample x by

$$\mathrm{var}(x) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}.$$

Example 23.6. If $x = (1, 3, -1)$, $\overline{x} = \frac{1+3-1}{3} = 1$, $x - \overline{x} = (0, 2, -2)$, and $\mathrm{var}(x) = \frac{0^2+2^2+(-2)^2}{2} = 4$. If $y = (1, 2, 3)$, $\overline{y} = \frac{1+2+3}{3} = 2$, $y - \overline{y} = (-1, 0, 1)$, and $\mathrm{var}(y) = \frac{(-1)^2+0^2+1^2}{2} = 2$.

There is a reason for using $n-1$ instead of n. The above definition makes $\mathrm{var}(x)$ an unbiased estimator of the variance of the random variable being sampled. However, we don't need to worry about this. Curious readers will find an explanation of these peculiar definitions in Epstein [57] (Chapter 14, Section 14.5) or in any decent statistics book.

Given two vectors $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$, the *sample covariance* (for short, *covariance*) of x and y is given by

$$\mathrm{cov}(x, y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n-1}.$$

Example 23.7. If we take $x = (1, 3, -1)$ and $y = (0, 2, -2)$, we know from Example 23.6 that $x - \overline{x} = (0, 2, -2)$ and $y - \overline{y} = (-1, 0, 1)$. Thus, $\mathrm{cov}(x, y) = \frac{0(-1)+2(0)+(-2)(1)}{2} = -1$.

The covariance of x and y measures how x and y vary from the mean with respect to each other. Obviously, $\mathrm{cov}(x, y) = \mathrm{cov}(y, x)$ and $\mathrm{cov}(x, x) = \mathrm{var}(x)$.

Note that

$$\mathrm{cov}(x, y) = \frac{(x - \overline{x})^\top (y - \overline{y})}{n-1}.$$

We say that x and y are *uncorrelated* iff $\mathrm{cov}(x, y) = 0$.

Finally, given an $n \times d$ matrix X of n points X_i, for PCA to be meaningful, it will be necessary to translate the origin to the *centroid* (or *center of gravity*) μ of the X_i's, defined by

$$\mu = \frac{1}{n}(X_1 + \cdots + X_n).$$

Observe that if $\mu = (\mu_1, \ldots, \mu_d)$, then μ_j is the mean of the vector C_j (the jth column of X).

We let $X - \mu$ denote the *matrix* whose ith row is the centered data point $X_i - \mu$ ($1 \leq i \leq n$). Then the *sample covariance matrix* (for short, *covariance matrix*) of X is the $d \times d$ symmetric matrix

$$\Sigma = \frac{1}{n-1}(X - \mu)^\top (X - \mu) = (\mathrm{cov}(C_i, C_j)).$$

Example 23.8. Let $X = \begin{pmatrix} 1 & 1 \\ 3 & 2 \\ -1 & 3 \end{pmatrix}$, the 3×2 matrix whose columns are the vector x and y of Example 23.6. Then

$$\mu = \frac{1}{3}\left[(1, 1) + (3, 2) + (-1, 3)\right] = (1, 2),$$

$$X - \mu = \begin{pmatrix} 0 & -1 \\ 2 & 0 \\ -2 & 1 \end{pmatrix},$$

and

$$\Sigma = \frac{1}{2}\begin{pmatrix} 0 & 2 & -2 \\ -1 & 0 & 1 \end{pmatrix}\begin{pmatrix} 0 & -1 \\ 2 & 0 \\ -2 & 1 \end{pmatrix} = \begin{pmatrix} 4 & -1 \\ -1 & 1 \end{pmatrix}.$$

Remark: The factor $\frac{1}{n-1}$ is irrelevant for our purposes and can be ignored.

Example 23.9. Here is the matrix $X - \mu$ in the case of our bearded mathematicians: since

$$\mu_1 = 1828.4, \quad \mu_2 = 5.6,$$

we get the following centered data set.

Name	year	length
Carl Friedrich Gauss	−51.4	−5.6
Camille Jordan	9.6	6.4
Adrien-Marie Legendre	−76.4	−5.6
Bernhard Riemann	−2.4	9.4
David Hilbert	33.6	−3.6
Henri Poincaré	25.6	−0.6
Emmy Noether	53.6	−5.6
Karl Weierstrass	13.4	−5.6
Eugenio Beltrami	6.6	−3.6
Hermann Schwarz	14.6	14.4

See Figure 23.3.

We can think of the vector C_j as representing the features of X in the direction e_j (the jth canonical basis vector in $\mathbb{R}^d$, namely $e_j = (0, \ldots, 1, \ldots 0)$, with a 1 in the jth position).

If $v \in \mathbb{R}^d$ is a unit vector, we wish to consider the projection of the data points $X_1, \ldots, X_n$ onto the line spanned by v. Recall from Euclidean geometry that if $x \in \mathbb{R}^d$ is any vector and $v \in \mathbb{R}^d$ is a unit vector, the projection of x onto the line spanned by v is

$$\langle x, v\rangle v.$$

Thus, with respect to the basis v, the projection of x has coordinate $\langle x, v\rangle$. If x is represented by a row vector and v by a column vector, then

$$\langle x, v\rangle = xv.$$

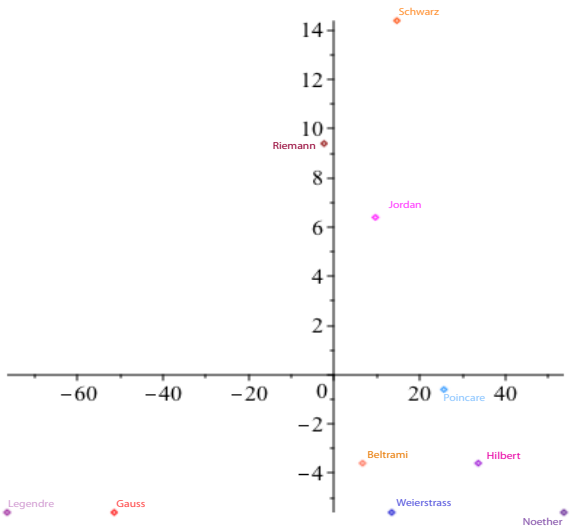

Figure 23.3: The centered data points of Example 23.9.

Therefore, the vector $Y \in \mathbb{R}^n$ consisting of the coordinates of the projections of $X_1, \ldots, X_n$ onto the line spanned by v is given by $Y = Xv$, and this is the linear combination

$$Xv = v_1 C_1 + \cdots + v_d C_d$$

of the columns of X (with $v = (v_1, \ldots, v_d)$).

Observe that because μ_j is the mean of the vector C_j (the jth column of X), we get

$$\overline{Y} = \overline{Xv} = v_1\mu_1 + \cdots + v_d\mu_d,$$

and so the centered point $Y - \overline{Y}$ is given by

$$Y - \overline{Y} = v_1(C_1 - \mu_1) + \cdots + v_d(C_d - \mu_d) = (X - \mu)v.$$

Furthermore, if $Y = Xv$ and $Z = Xw$, then

$$\begin{aligned} \operatorname{cov}(Y, Z) &= \frac{((X - \mu)v)^\top (X - \mu)w}{n - 1} \\ &= v^\top \frac{1}{n-1}(X - \mu)^\top (X - \mu)w \\ &= v^\top \Sigma w, \end{aligned}$$

where Σ is the covariance matrix of X. Since $Y - \overline{Y}$ has zero mean, we have

$$\operatorname{var}(Y) = \operatorname{var}(Y - \overline{Y}) = v^\top \frac{1}{n-1}(X - \mu)^\top (X - \mu)v.$$

The above suggests that we should move the origin to the centroid μ of the X_i's and consider the matrix $X - \mu$ of the centered data points $X_i - \mu$.

From now on beware that we denote the columns of $X - \mu$ by $C_1, \ldots, C_d$ and that Y denotes the *centered* point $Y = (X - \mu)v = \sum_{j=1}^{d} v_j C_j$, where v is a unit vector.

Basic idea of PCA: The principal components of X are *uncorrelated* projections Y of the data points X_1, ..., X_n onto some directions v (where the v's are unit vectors) such that $\mathrm{var}(Y)$ is maximal.

This suggests the following definition:

Definition 23.2. Given an $n \times d$ matrix X of data points $X_1, \ldots, X_n$, if μ is the centroid of the X_i's, then a *first principal component of X (first PC)* is a centered point $Y_1 = (X - \mu)v_1$, the projection of $X_1, \ldots, X_n$ onto a direction v_1 such that $\mathrm{var}(Y_1)$ is maximized, where v_1 is a unit vector (recall that $Y_1 = (X - \mu)v_1$ is a linear combination of the C_j's, the columns of $X - \mu$).

More generally, if $Y_1, \ldots, Y_k$ are k principal components of X along some unit vectors $v_1, \ldots, v_k$, where $1 \leq k < d$, a *$(k+1)$th principal component of X ($(k+1)$th PC)* is a centered point $Y_{k+1} = (X - \mu)v_{k+1}$, the projection of $X_1, \ldots, X_n$ onto some direction v_{k+1} such that $\mathrm{var}(Y_{k+1})$ is maximized, subject to $\mathrm{cov}(Y_h, Y_{k+1}) = 0$ for all h with $1 \leq h \leq k$, and where v_{k+1} is a unit vector (recall that $Y_h = (X - \mu)v_h$ is a linear combination of the C_j's). The v_h are called *principal directions*.

The following proposition is the key to the main result about PCA. This result was already proven in Proposition 17.23 except that the eigenvalues were listed in increasing order. For the reader's convenience we prove it again.

Proposition 23.10. *If A is a symmetric $d \times d$ matrix with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ and if $(u_1, \ldots, u_d)$ is any orthonormal basis of eigenvectors of A, where u_i is a unit eigenvector associated with λ_i, then*

$$\max_{x \neq 0} \frac{x^\top A x}{x^\top x} = \lambda_1$$

(with the maximum attained for $x = u_1$) and

$$\max_{x \neq 0, x \in \{u_1, \ldots, u_k\}^\perp} \frac{x^\top A x}{x^\top x} = \lambda_{k+1}$$

(with the maximum attained for $x = u_{k+1}$), where $1 \leq k \leq d - 1$.

Proof. First observe that

$$\max_{x \neq 0} \frac{x^\top A x}{x^\top x} = \max_{x} \{x^\top A x \mid x^\top x = 1\},$$

and similarly,

$$\max_{x\neq 0, x\in\{u_1,\ldots,u_k\}^\perp} \frac{x^\top Ax}{x^\top x} = \max_x \left\{x^\top Ax \mid (x \in \{u_1,\ldots,u_k\}^\perp) \wedge (x^\top x = 1)\right\}.$$

Since A is a symmetric matrix, its eigenvalues are real and it can be diagonalized with respect to an orthonormal basis of eigenvectors, so let $(u_1,\ldots,u_d)$ be such a basis. If we write

$$x = \sum_{i=1}^d x_i u_i,$$

a simple computation shows that

$$x^\top Ax = \sum_{i=1}^d \lambda_i x_i^2.$$

If $x^\top x = 1$, then $\sum_{i=1}^d x_i^2 = 1$, and since we assumed that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$, we get

$$x^\top Ax = \sum_{i=1}^d \lambda_i x_i^2 \leq \lambda_1\left(\sum_{i=1}^d x_i^2\right) = \lambda_1.$$

Thus,

$$\max_x \left\{x^\top Ax \mid x^\top x = 1\right\} \leq \lambda_1,$$

and since this maximum is achieved for $e_1 = (1, 0, \ldots, 0)$, we conclude that

$$\max_x \left\{x^\top Ax \mid x^\top x = 1\right\} = \lambda_1.$$

Next observe that $x \in \{u_1,\ldots,u_k\}^\perp$ and $x^\top x = 1$ iff $x_1 = \cdots = x_k = 0$ and $\sum_{i=1}^d x_i = 1$. Consequently, for such an x, we have

$$x^\top Ax = \sum_{i=k+1}^d \lambda_i x_i^2 \leq \lambda_{k+1}\left(\sum_{i=k+1}^d x_i^2\right) = \lambda_{k+1}.$$

Thus,

$$\max_x \left\{x^\top Ax \mid (x \in \{u_1,\ldots,u_k\}^\perp) \wedge (x^\top x = 1)\right\} \leq \lambda_{k+1},$$

and since this maximum is achieved for $e_{k+1} = (0,\ldots,0,1,0,\ldots,0)$ with a 1 in position $k+1$, we conclude that

$$\max_x \left\{x^\top Ax \mid (x \in \{u_1,\ldots,u_k\}^\perp) \wedge (x^\top x = 1)\right\} = \lambda_{k+1},$$

as claimed. □

The quantity

$$\frac{x^\top Ax}{x^\top x}$$

is known as the *Rayleigh ratio* or *Rayleigh–Ritz ratio* (see Section 17.6) and Proposition 23.10 is often known as part of the *Rayleigh–Ritz theorem*.

Proposition 23.10 also holds if A is a Hermitian matrix and if we replace $x^\top Ax$ by x^*Ax and $x^\top x$ by x^*x. The proof is unchanged, since a Hermitian matrix has real eigenvalues and is diagonalized with respect to an orthonormal basis of eigenvectors (with respect to the Hermitian inner product).

We then have the following fundamental result showing how *the SVD of X yields the PCs*:

Theorem 23.11. *(SVD yields PCA) Let X be an $n \times d$ matrix of data points $X_1, \ldots, X_n$, and let μ be the centroid of the X_i's. If $X - \mu = VDU^\top$ is an SVD decomposition of $X - \mu$ and if the main diagonal of D consists of the singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d$, then the centered points $Y_1, \ldots, Y_d$, where*

$$Y_k = (X - \mu)u_k = k\text{th column of } VD$$

and u_k is the kth column of U, are d principal components of X. Furthermore,

$$\mathrm{var}(Y_k) = \frac{\sigma_k^2}{n-1}$$

and $\mathrm{cov}(Y_h, Y_k) = 0$, whenever $h \neq k$ and $1 \leq k, h \leq d$.

Proof. Recall that for any unit vector v, the centered projection of the points $X_1, \ldots, X_n$ onto the line of direction v is $Y = (X - \mu)v$ and that the variance of Y is given by

$$\mathrm{var}(Y) = v^\top \frac{1}{n-1}(X - \mu)^\top (X - \mu)v.$$

Since $X - \mu = VDU^\top$, we get

$$\begin{aligned}
\mathrm{var}(Y) &= v^\top \frac{1}{(n-1)}(X - \mu)^\top (X - \mu)v \\
&= v^\top \frac{1}{(n-1)} UDV^\top VDU^\top v \\
&= v^\top U \frac{1}{(n-1)} D^2 U^\top v.
\end{aligned}$$

Similarly, if $Y = (X - \mu)v$ and $Z = (X - \mu)w$, then the covariance of Y and Z is given by

$$\mathrm{cov}(Y, Z) = v^\top U \frac{1}{(n-1)} D^2 U^\top w.$$

Obviously, $U\frac{1}{(n-1)}D^2U^\top$ is a symmetric matrix whose eigenvalues are $\frac{\sigma_1^2}{n-1} \geq \cdots \geq \frac{\sigma_d^2}{n-1}$, and the columns of U form an orthonormal basis of unit eigenvectors.

We proceed by induction on k. For the base case, $k = 1$, maximizing $\mathrm{var}(Y)$ is equivalent to maximizing

$$v^\top U\frac{1}{(n-1)}D^2U^\top v,$$

where v is a unit vector. By Proposition 23.10, the maximum of the above quantity is the largest eigenvalue of $U\frac{1}{(n-1)}D^2U^\top$, namely $\frac{\sigma_1^2}{n-1}$, and it is achieved for u_1, the first columnn of U. Now we get

$$Y_1 = (X - \mu)u_1 = VDU^\top u_1,$$

and since the columns of U form an orthonormal basis, $U^\top u_1 = e_1 = (1, 0, \ldots, 0)$, and so Y_1 is indeed the first column of VD.

By the induction hypothesis, the centered points $Y_1, \ldots, Y_k$, where $Y_h = (X - \mu)u_h$ and $u_1, \ldots, u_k$ are the first k columns of U, are k principal components of X. Because

$$\mathrm{cov}(Y, Z) = v^\top U\frac{1}{(n-1)}D^2U^\top w,$$

where $Y = (X - \mu)v$ and $Z = (X - \mu)w$, the condition $\mathrm{cov}(Y_h, Z) = 0$ for $h = 1, \ldots, k$ is equivalent to the fact that w belongs to the orthogonal complement of the subspace spanned by $\{u_1, \ldots, u_k\}$, and maximizing $\mathrm{var}(Z)$ subject to $\mathrm{cov}(Y_h, Z) = 0$ for $h = 1, \ldots, k$ is equivalent to maximizing

$$w^\top U\frac{1}{(n-1)}D^2U^\top w,$$

where w is a unit vector orthogonal to the subspace spanned by $\{u_1, \ldots, u_k\}$. By Proposition 23.10, the maximum of the above quantity is the $(k+1)$th eigenvalue of $U\frac{1}{(n-1)}D^2U^\top$, namely $\frac{\sigma_{k+1}^2}{n-1}$, and it is achieved for u_{k+1}, the $(k+1)$th columnn of U. Now we get

$$Y_{k+1} = (X - \mu)u_{k+1} = VDU^\top u_{k+1},$$

and since the columns of U form an orthonormal basis, $U^\top u_{k+1} = e_{k+1}$, and Y_{k+1} is indeed the $(k+1)$th column of VD, which completes the proof of the induction step. □

The d columns $u_1, \ldots, u_d$ of U are usually called the *principal directions* of $X - \mu$ (and X). We note that not only do we have $\mathrm{cov}(Y_h, Y_k) = 0$ whenever $h \neq k$, but the directions $u_1, \ldots, u_d$ along which the data are projected are mutually orthogonal.

Example 23.10. For the centered data set of our bearded mathematicians (Example 23.9) we have $X - \mu = V\Sigma U^\top$, where Σ has two nonzero singular values, $\sigma_1 = 116.9803, \sigma_2 = 21.7812$, and with

$$U = \begin{pmatrix} 0.9995 & 0.0325 \\ 0.0325 & -0.9995 \end{pmatrix},$$

so the principal directions are $u_1 = (0.9995, 0.0325)$ and $u_2 = (0.0325, -0.9995)$. Observe that u_1 is almost the direction of the x-axis, and u_2 is almost the opposite direction of the y-axis. We also find that the projections Y_1 and Y_2 along the principal directions are

$$VD = \begin{pmatrix} -51.5550 & 3.9249 \\ 9.8031 & -6.0843 \\ -76.5417 & 3.1116 \\ -2.0929 & -9.4731 \\ 33.4651 & 4.6912 \\ 25.5669 & 1.4325 \\ 53.3894 & 7.3408 \\ 13.2107 & 6.0330 \\ 6.4794 & 3.8128 \\ 15.0607 & -13.9174 \end{pmatrix}, \quad \text{with} \quad X - \mu = \begin{pmatrix} -51.4000 & -5.6000 \\ 9.6000 & 6.4000 \\ -76.4000 & -5.6000 \\ -2.4000 & 9.4000 \\ 33.6000 & -3.6000 \\ 25.6000 & -0.6000 \\ 53.6000 & -5.6000 \\ 13.4000 & -5.6000 \\ 6.6000 & -3.6000 \\ 14.6000 & 14.4000 \end{pmatrix}.$$

See Figures 23.4, 23.5, and 23.6.

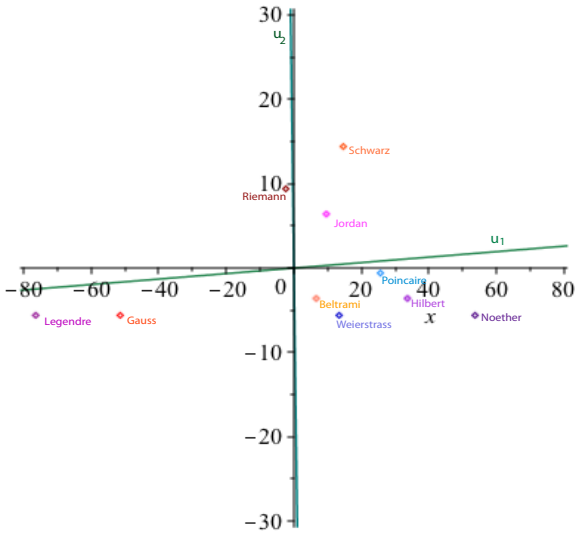

Figure 23.4: The centered data points of Example 23.9 and the two principal directions of Example 23.10.

We know from our study of SVD that $\sigma_1^2, \ldots, \sigma_d^2$ are the eigenvalues of the symmetric positive semidefinite matrix $(X - \mu)^\top (X - \mu)$ and that $u_1, \ldots, u_d$ are corresponding eigenvectors. Numerically, it is preferable to use SVD on $X - \mu$ rather than to compute explicitly $(X - \mu)^\top (X - \mu)$ and then diagonalize it. Indeed, the explicit computation of $A^\top A$ from a matrix A can be numerically quite unstable, and good SVD algorithms avoid computing $A^\top A$ explicitly.

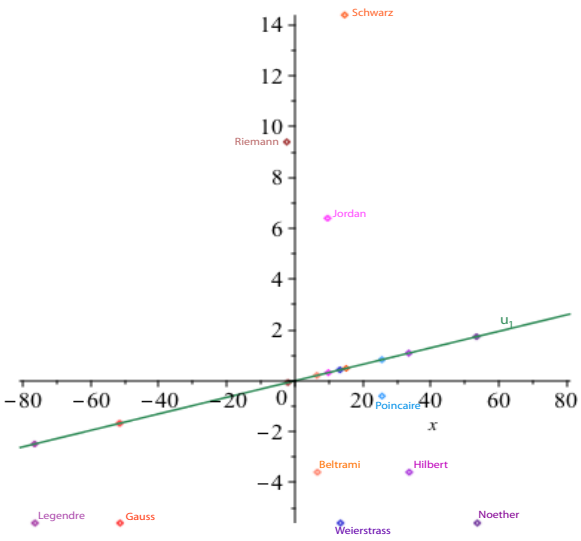

Figure 23.5: The first principal components of Example 23.10, i.e. the projection of the centered data points onto the u_1 line.

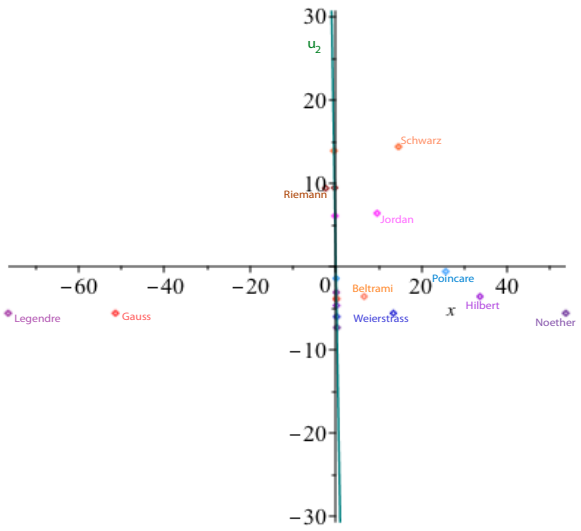

Figure 23.6: The second principal components of Example 23.10, i.e. the projection of the centered data points onto the u_2 line.

In general, since an SVD of X is not unique, *the principal directions $u_1, \ldots, u_d$ are not unique*. This can happen when a data set has some *rotational symmetries*, and in such a case, PCA is not a very good method for analyzing the data set.

23.5 Best Affine Approximation

A problem very close to PCA (and based on least squares) is to *best approximate a data set of n points $X_1, \ldots, X_n$, with $X_i \in \mathbb{R}^d$, by a p-dimensional affine subspace A of $\mathbb{R}^d$, with* $1 \leq p \leq d-1$ (the terminology rank $d-p$ is also used).

First consider $p = d-1$. Then $A = A_1$ is an affine hyperplane (in $\mathbb{R}^d$), and it is given by an equation of the form

$$a_1 x_1 + \cdots + a_d x_d + c = 0.$$

By *best approximation*, we mean that $(a_1, \ldots, a_d, c)$ solves the homogeneous linear system

$$\begin{pmatrix} x_{1\,1} & \cdots & x_{1\,d} & 1 \\ \vdots & \vdots & \vdots & \vdots \\ x_{n\,1} & \cdots & x_{n\,d} & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ \vdots \\ a_d \\ c \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}$$

in the *least squares sense, subject to the condition that $a = (a_1, \ldots, a_d)$ is a unit vector*, that is, $a^\top a = 1$, where $X_i = (x_{i\,1}, \cdots, x_{i\,d})$.

If we form the symmetric matrix

$$\begin{pmatrix} x_{1\,1} & \cdots & x_{1\,d} & 1 \\ \vdots & \vdots & \vdots & \vdots \\ x_{n\,1} & \cdots & x_{n\,d} & 1 \end{pmatrix}^\top \begin{pmatrix} x_{1\,1} & \cdots & x_{1\,d} & 1 \\ \vdots & \vdots & \vdots & \vdots \\ x_{n\,1} & \cdots & x_{n\,d} & 1 \end{pmatrix}$$

involved in the normal equations, we see that the bottom row (and last column) of that matrix is

$$n\mu_1 \quad \cdots \quad n\mu_d \quad n,$$

where $n\mu_j = \sum_{i=1}^n x_{i\,j}$ is n times the mean of the column C_j of X.

Therefore, if $(a_1, \ldots, a_d, c)$ is a least squares solution, that is, a solution of the normal equations, we must have

$$n\mu_1 a_1 + \cdots + n\mu_d a_d + nc = 0,$$

that is,

$$a_1\mu_1 + \cdots + a_d\mu_d + c = 0,$$

which means that the *hyperplane A_1 must pass through the centroid μ of the data points* $X_1, \ldots, X_n$. Then we can rewrite the original system with respect to the centered data $X_i - \mu$, find that the variable c drops out, get the system

$$(X - \mu)a = 0,$$

where $a = (a_1, \ldots, a_d)$.

Thus, we are looking for a unit vector a solving $(X - \mu)a = 0$ in the least squares sense, that is, some a such that $a^\top a = 1$ minimizing

$$a^\top (X - \mu)^\top (X - \mu) a.$$

Compute some SVD $VDU^\top$ of $X - \mu$, where the main diagonal of D consists of the singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d$ of $X - \mu$ arranged in descending order. Then

$$a^\top (X - \mu)^\top (X - \mu) a = a^\top U D^2 U^\top a,$$

where $D^2 = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$ is a diagonal matrix, so pick a to be *the last column in* U (corresponding to the smallest eigenvalue σ_d^2 of $(X - \mu)^\top (X - \mu)$). This is a solution to our best fit problem.

Therefore, if U_{d-1} is the linear hyperplane defined by a, that is,

$$U_{d-1} = \{u \in \mathbb{R}^d \mid \langle u, a \rangle = 0\},$$

where a is the last column in U for some SVD $VDU^\top$ of $X - \mu$, we have shown that the affine hyperplane $A_1 = \mu + U_{d-1}$ is a best approximation of the data set $X_1, \ldots, X_n$ in the least squares sense.

It is easy to show that this hyperplane $A_1 = \mu + U_{d-1}$ minimizes the sum of the square distances of each X_i to its orthogonal projection onto A_1. Also, since U_{d-1} is the orthogonal complement of a, the last column of U, we see that U_{d-1} is spanned by the first $d-1$ columns of U, that is, the first $d-1$ principal directions of $X - \mu$.

All this can be generalized to a *best* $(d-k)$*-dimensional affine subspace* A_k *approximating* $X_1, \ldots, X_n$ *in the least squares sense* $(1 \leq k \leq d-1)$. Such an affine subspace A_k is cut out by k independent hyperplanes H_i (with $1 \leq i \leq k$), each given by some equation

$$a_{i\,1} x_1 + \cdots + a_{i\,d} x_d + c_i = 0.$$

If we write $a_i = (a_{i\,1}, \cdots, a_{i\,d})$, to say that the H_i are independent means that $a_1, \ldots, a_k$ are linearly independent. In fact, we may assume that $a_1, \ldots, a_k$ form an *orthonormal system*.

Then finding a best $(d-k)$-dimensional affine subspace A_k amounts to solving the homogeneous linear system

$$\begin{pmatrix} X & \mathbf{1} & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & X & \mathbf{1} \end{pmatrix} \begin{pmatrix} a_1 \\ c_1 \\ \vdots \\ a_k \\ c_k \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix},$$

in the least squares sense, subject to the conditions $a_i^\top a_j = \delta_{i\,j}$, for all i, j with $1 \leq i, j \leq k$, where the matrix of the system is a block diagonal matrix consisting of k diagonal blocks $(X, \mathbf{1})$, where $\mathbf{1}$ denotes the column vector $(1, \ldots, 1) \in \mathbb{R}^n$.

Again it is easy to see that each hyperplane H_i *must pass through the centroid* μ of $X_1, \ldots, X_n$, and by switching to the centered data $X_i - \mu$ we get the system

$$\begin{pmatrix} X - \mu & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & X - \mu \end{pmatrix} \begin{pmatrix} a_1 \\ \vdots \\ a_k \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix},$$

with $a_i^\top a_j = \delta_{i\,j}$ for all i, j with $1 \leq i, j \leq k$.

If $VDU^\top = X - \mu$ is an SVD decomposition, it is easy to see that a least squares solution of this system is given by the last k columns of U, assuming that the main diagonal of D consists of the singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d$ of $X - \mu$ arranged in descending order. But now the $(d-k)$-dimensional subspace U_{d-k} cut out by the hyperplanes defined by $a_1, \ldots, a_k$ is simply the orthogonal complement of $(a_1, \ldots, a_k)$, which is the subspace spanned by the first $d - k$ columns of U.

So the best $(d-k)$-dimensional affine subpsace A_k approximating $X_1, \ldots, X_n$ in the least squares sense is

$$A_k = \mu + U_{d-k},$$

where U_{d-k} is the linear subspace spanned by the first $d-k$ principal directions of $X - \mu$, that is, the first $d-k$ columns of U. Consequently, we get the following interesting interpretation of PCA (actually, principal directions):

Theorem 23.12. *Let X be an $n \times d$ matrix of data points $X_1, \ldots, X_n$, and let μ be the centroid of the X_i's. If $X - \mu = VDU^\top$ is an SVD decomposition of $X - \mu$ and if the main diagonal of D consists of the singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d$, then a best $(d-k)$-dimensional affine approximation A_k of $X_1, \ldots, X_n$ in the least squares sense is given by*

$$A_k = \mu + U_{d-k},$$

where U_{d-k} is the linear subspace spanned by the first $d - k$ columns of U, the first $d - k$ principal directions of $X - \mu$ ($1 \leq k \leq d - 1$).

Example 23.11. Going back to Example 23.10, a best 1-dimensional affine approximation A_1 is the affine line passing through $(\mu_1, \mu_2) = (1824.4, 5.6)$ of direction $u_1 = (0.9995, 0.0325)$.

Example 23.12. Suppose in the data set of Example 23.5 that we add the month of birth of every mathematician as a feature. We obtain the following data set.

Name	month	year	length
Carl Friedrich Gauss	4	1777	0
Camille Jordan	1	1838	12
Adrien-Marie Legendre	9	1752	0
Bernhard Riemann	9	1826	15
David Hilbert	1	1862	2
Henri Poincaré	4	1854	5
Emmy Noether	3	1882	0
Karl Weierstrass	10	1815	0
Eugenio Beltrami	10	1835	2
Hermann Schwarz	1	1843	20

The mean of the first column is 5.2, and the centered data set is given below.

Name	month	year	length
Carl Friedrich Gauss	−1.2	−51.4	−5.6
Camille Jordan	−4.2	9.6	6.4
Adrien-Marie Legendre	3.8	−76.4	−5.6
Bernhard Riemann	3.8	−2.4	9.4
David Hilbert	−4.2	33.6	−3.6
Henri Poincaré	−1.2	25.6	−0.6
Emmy Noether	−2.2	53.6	−5.6
Karl Weierstrass	4.8	13.4	−5.6
Eugenio Beltrami	4.8	6.6	−3.6
Hermann Schwarz	−4.2	14.6	14.4

Running SVD on this data set we get

$$U = \begin{pmatrix} 0.0394 & 0.1717 & 0.9844 \\ -0.9987 & 0.0390 & 0.0332 \\ -0.0327 & -0.9844 & 0.1730 \end{pmatrix}, \qquad D = \begin{pmatrix} 117.0706 & 0 & 0 \\ 0 & 22.0390 & 0 \\ 0 & 0 & 10.1571 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

and

$$VD = \begin{pmatrix} 51.4683 & 3.3013 & -3.8569 \\ -9.9623 & -6.6467 & -2.7082 \\ 76.6327 & 3.1845 & 0.2348 \\ 2.2393 & -8.6943 & 5.2872 \\ -33.6038 & 4.1334 & -3.6415 \\ -25.5941 & 1.3833 & -0.4350 \\ -53.4333 & 7.2258 & -1.3547 \\ -13.0100 & 6.8594 & 4.2010 \\ -6.2843 & 4.6254 & 4.3212 \\ -15.2173 & -14.3266 & -1.1581 \end{pmatrix}, \quad X - \mu = \begin{pmatrix} -1.2000 & -51.4000 & -5.6000 \\ -4.2000 & 9.6000 & 6.4000 \\ 3.8000 & -76.4000 & -5.6000 \\ 3.8000 & -2.4000 & 9.4000 \\ -4.2000 & 33.6000 & -3.6000 \\ -1.2000 & 25.6000 & -0.6000 \\ -2.2000 & 53.6000 & -5.6000 \\ 4.8000 & 13.4000 & -5.6000 \\ 4.8000 & 6.6000 & -3.6000 \\ -4.2000 & 14.6000 & 14.4000 \end{pmatrix}.$$

The first principal direction $u_1 = (0.0394, -0.9987, -0.0327)$ is basically the opposite of the y-axis, and the most significant feature is the year of birth. The second principal direction $u_2 = (0.1717, 0.0390, -0.9844)$ is close to the opposite of the z-axis, and the second most significant feature is the lenght of beards. A best affine plane is spanned by the vectors u_1 and u_2.

There are many applications of PCA to data compression, dimension reduction, and pattern analysis. The basic idea is that in many cases, given a data set $X_1, \ldots, X_n$, with $X_i \in \mathbb{R}^d$, only a "small" subset of $m < d$ of the features is needed to describe the data set accurately.

If $u_1, \ldots, u_d$ are the principal directions of $X - \mu$, then the first m projections of the data (the first m principal components, i.e., the first m columns of VD) onto the first m principal directions represent the data without much loss of information. Thus, instead of using the original data points $X_1, \ldots, X_n$, with $X_i \in \mathbb{R}^d$, we can use their projections onto the first m principal directions $Y_1, \ldots, Y_m$, where $Y_i \in \mathbb{R}^m$ and $m < d$, obtaining a compressed version of the original data set.

For example, PCA is used in computer vision for *face recognition*. Sirovitch and Kirby (1987) seem to be the first to have had the idea of using PCA to compress facial images. They introduced the term *eigenpicture* to refer to the principal directions, u_i. However, an explicit face recognition algorithm was given only later by Turk and Pentland (1991). They renamed eigenpictures as *eigenfaces*.

For details on the topic of eigenfaces, see Forsyth and Ponce [64] (Chapter 22, Section 22.3.2), where you will also find exact references to Turk and Pentland's papers.

Another interesting application of PCA is to the *recognition of handwritten digits*. Such an application is described in Hastie, Tibshirani, and Friedman, [88] (Chapter 14, Section 14.5.1).

23.6 Summary

The main concepts and results of this chapter are listed below:

- *Least squares problems*.
- Existence of a least squares solution of smallest norm (Theorem 23.1).
- The *pseudo-inverse* A^+ of a matrix A.
- The least squares solution of smallest norm is given by the pseudo-inverse (Theorem 23.2)
- Projection properties of the pseudo-inverse.
- The pseudo-inverse of a normal matrix.
- The *Penrose characterization* of the pseudo-inverse.
- Data compression and SVD.
- Best approximation of rank $< r$ of a matrix.
- *Principal component analysis*.
- Review of basic statistical concepts: *mean, variance, covariance*, *covariance matrix*.
- Centered data, *centroid*.
- The *principal components* (*PCA*).
- The *Rayleigh–Ritz theorem* (Theorem 23.10).
- The main theorem: *SVD yields PCA* (Theorem 23.11).
- Best affine approximation.
- SVD yields a best affine approximation (Theorem 23.12).
- Face recognition, eigenfaces.

23.7 Problems

Problem 23.1. Consider the overdetermined system in the single variable x:

$$a_1 x = b_1, \ldots, a_m x = b_m,$$

with $a_1^2 + \cdots + a_m^2 \neq 0$. Prove that the least squares solution of smallest norm is given by

$$x^+ = \frac{a_1 b_1 + \cdots + a_m b_m}{a_1^2 + \cdots + a_m^2}.$$