Chapter 13

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 à faire connaitre l'état du système le plus proche de la vérité.
>
> —**Legendre, 1805,** *Nouvelles Méthodes pour la détermination des Orbites des Comètes*

13.1 Least Squares Problems and the Pseudo-inverse

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.

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 $y = dx + c$.

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

$$\begin{aligned} c + dx_1 &= y_1, \\ c + dx_2 &= y_2, \\ c + dx_3 &= y_3. \end{aligned}$$

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,

$$(c + dx_1 - y_1)^2 + (c + dx_2 - y_2)^2 + (c + dx_3 - y_3)^2.$$

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

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

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, the normal equations for the above problem are

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

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

The proof also shows that x minimizes $\|Ax - b\|_2^2$ iff

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

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 13.1. Given any $m \times n$ matrix A, if $A = VDU^\top$ is an SVD of A with

$$D = \text{diag}(\lambda_1, \ldots, \lambda_r, 0, \ldots, 0),$$

where D is an $m \times n$ matrix and $\lambda_i > 0$, if we let

$$D^+ = \text{diag}(1/\lambda_1, \ldots, 1/\lambda_r, 0, \ldots, 0),$$

an $n \times m$ matrix, the *pseudo-inverse* of A is defined by

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

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

$$x^+ = A^+ b = U D^+ V^\top b.$$

By Proposition 13.2 and Theorem 13.1, $A^+ b$ is uniquely defined by every b, and thus A^+ depends only on A.

Let $A = U \Sigma V^\top$ be an SVD for A. It is easy to check that

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

and both AA^+ and A^+A are symmetric matrices.

In fact,

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

and

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

We immediately get

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

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

We claim that 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* $A^\top$.

It is also useful to know that $\text{range}(A) = \text{range}(AA^+)$ consists of all vectors $y \in \mathbb{R}^n$ such that

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

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

Similarly, $\text{range}(A^+A) = \text{Ker}(A)^\perp$ consists of all vectors $y \in \mathbb{R}^n$ such that

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

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

If A is a symmetric matrix, then in general, there is no SVD $U\Sigma V^\top$ of A with $U = V$.

However, if A is positive semidefinite, then the eigenvalues of A are nonnegative, and so the nonzero eigenvalues of A are equal to the singular values of A and SVDs of A are of the form

$$A = U\Sigma U^\top.$$

Analogous results hold for complex matrices, but in this case, U and V 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.

If A is a (real) normal matrix, then we know from Theorem 10.15 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

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

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 $B_k = (\lambda_k)$.

Proposition 13.3. *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

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

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 [20].

Proposition 13.4. *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:*

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

13.2 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 4 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, $k < r$) so that $\|A - B\|_2$ (or $\|A - B\|_F$) is minimized.

Proposition 13.5. *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*

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

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

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.

A nice example of the use of Proposition 13.5 in image compression is given in Demmel [11], Chapter 3, Section 3.2.3, pages 113–115; see the Matlab demo.

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 $A^\top A$).

Interested readers should read Section 5.4 of Demmel's excellent book [11], which contains an overview of most known methods and an extensive list of references.

13.3 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 $X_i = (x_{i\,1}, \ldots, x_{i\,d})$, each $x_{i\,j}$ is the value of some *feature* (or *attribute*) of that person.

For example, the X_i's could be mathematicians, $d = 2$, and the first component, $x_{i\,1}$, of X_i could be the year that X_i was born, and the second component, $x_{i\,2}$, 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.

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

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

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.

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

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

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

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

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

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

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

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 $\mathbb{R}^d$).

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

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

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), the centered point $Y - \overline{Y}$ is given by

$$Y - \overline{Y} = (X - \mu)v.$$

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

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

where

$$\Sigma = \frac{1}{n - 1}(X - \mu)^\top (X - \mu)$$

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

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

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 13.2. 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:

Proposition 13.6. *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 $1 \leq k \leq d - 1$.

The quantity

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

is known as the *Rayleigh–Ritz ratio* and Proposition 13.6 is often known as part of the *Rayleigh–Ritz theorem*.

Proposition 13.6 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.

Theorem 13.7. *(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*

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

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 $1 \leq k, h \leq d$.

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.

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.

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.

13.4 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 $X_i - \mu$, and we find that the variable c drops out and we 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.

Is 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.

Theorem 13.8. *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 $X - \mu$ $(1 \leq k \leq d-1)$.

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*.

Another interesting application of PCA is to the *recognition of handwritten digits*. Such an application is described in Hastie, Tibshirani, and Friedman, [17] (Chapter 14, Section 14.5.1).