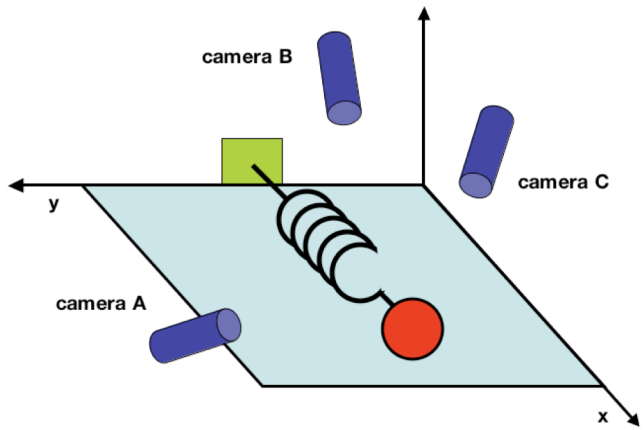
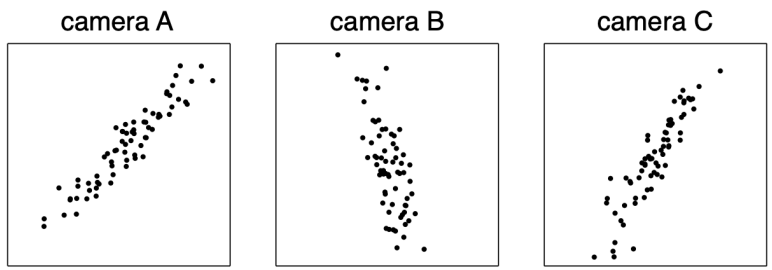
13: Principal Component Analysis

"*a statistical interpretation of the singular value decomposition*"

- motivating examples
- covariance matrices
- PCA

Example 1: An ignorant experimentalist





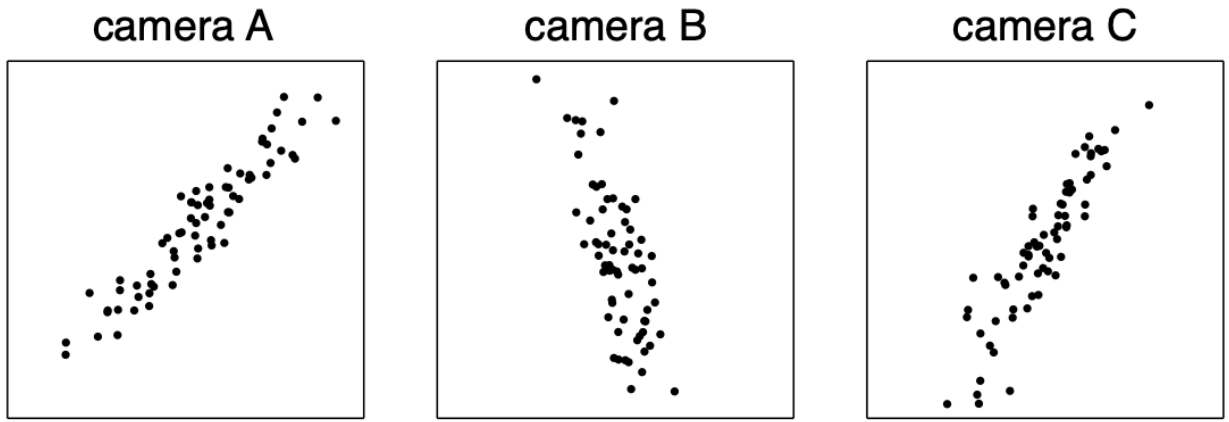
- 3 cameras recording at 120Hz
- each camera records an image indicating a two dimen-sional position of the ball
- arbitrary camera positions

[**A Tutorial on Principal Component Analysis**, Jonathon Shlens, 2014]

Goal

determine that $\mathbf{x}$ i.e. the unit basis vector along the x-axis, is the important dimension

Data



- 120Hz recording for 10 mins: $120 \times 10 \times 60 = 72000$ measurements of the form

$$X^t = \begin{bmatrix} x_A^t \\ y_A^t \\ x_B^t \\ y_B^t \\ x_C^t \\ y_C^t \end{bmatrix} \qquad X = \begin{bmatrix} X^1 & X^2 & \dots & X^{72000} \end{bmatrix}$$

Some linear algebra

- each sample X^t is a vector in $\mathbb{R}^m$, where m denotes no. of measurement types
- equivalently, every sample is a vector that lies in an m-dimensional vector
- a vector space is spanned by some orthonormal basis

Some linear algebra

- each sample X^t is a vector in $\mathbb{R}^m$, where m denotes no. of measurement types
- equivalently, every sample is a vector that lies in an m-dimensional vector
- a vector space is spanned by some orthonormal basis
- which orthonormal basis best represents the structure of our measurement data?

Noise

further complicated by the fact that data contains noise

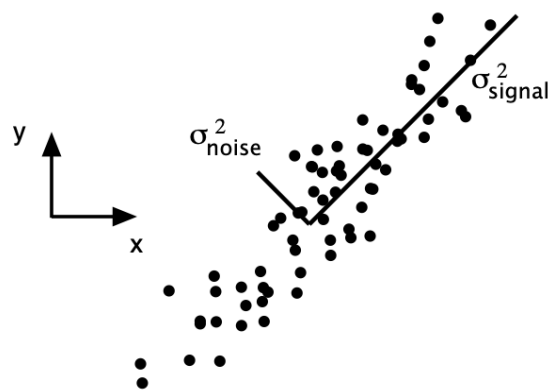
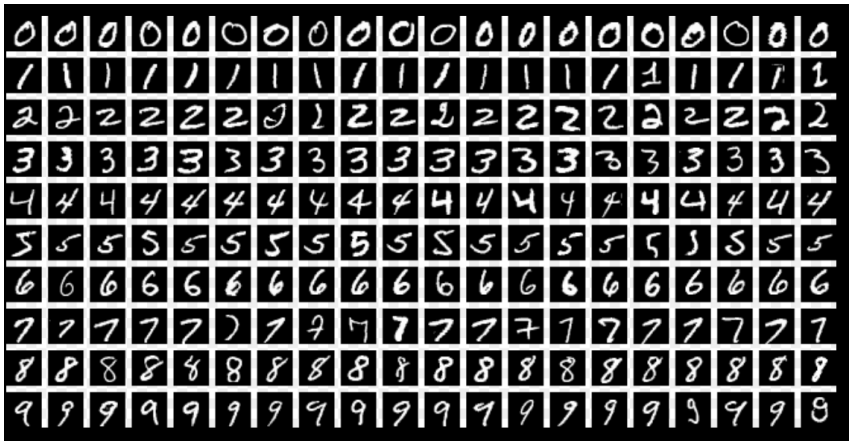


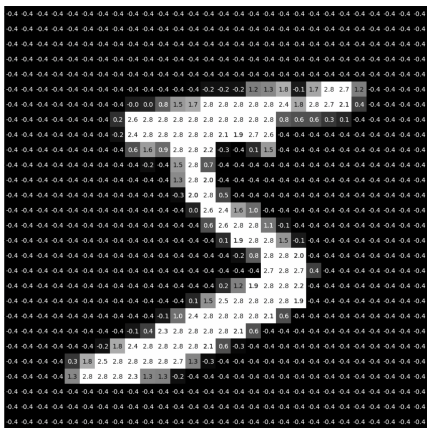
Figure: Camera A measurements

Example 2: Low dimensional structure

recall the MNIST data set

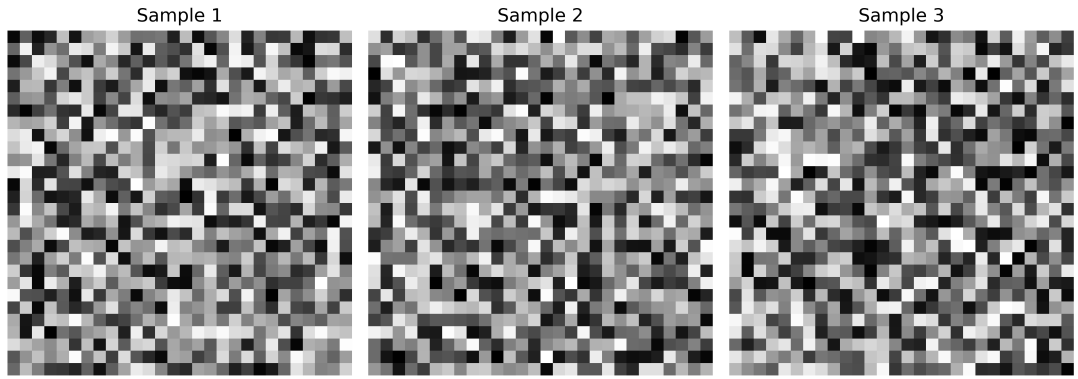


and the random sample



which we view as a 784 dimensional vector (28×28 pixels)

$\mathbb{R}^{784}$ is a vector space and MNIST images live inside this space



but most elements of $\mathbb{R}^{754}$ are **not** MNIST characters

Key Observation

Although MNIST images are vectors in a 784-dimensional space, the actual set of handwritten digits occupies only a tiny portion of that space. **Most** of the **variation** among digits can be captured by a much **lower-dimensional** structure.

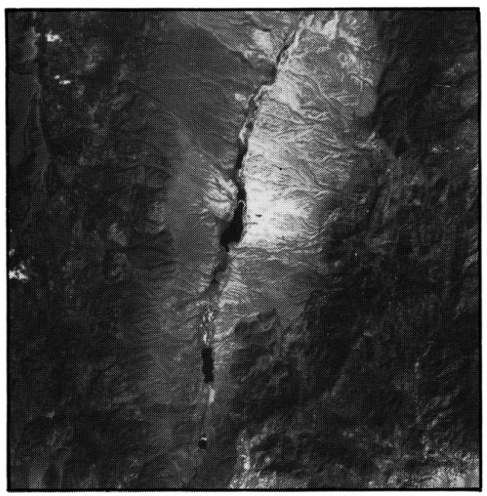
Principal Component Analysis

PCA is a technique that finds the best low-dimensional subspace that approximates this structure

Example 3: Multichannel Image Processing

Landsat satellite images

two Landsat satellites traveling in near polar orbits, record images of the earth in swaths 185 Km wide



(a) Spectral band 1: Visible blue.

(b) Spectral band 4: Near infrared.

(c) Spectral band 7: Mid-infrared.

- each satellite records 7 simultaneous images
  - three in the visible light spectrum
  - four in infrared and thermal bands
- images are digitized and stored as a matrix
- each entry indicates the signal intensity at a corresponding small point (or pixel)

By the numbers

- two satellites, orbiting the earth one every 80mins
- every 16 days, each satellite passes over almost every square kilometer of the earth's surface
- an image covers an area approximately $185\mathrm{Km} \times 185\mathrm{Km}$
- each images contain $2,000 \times 2,000 = 4,000,000$ pixels
- 7 images per $185\mathrm{Km} \times 185\mathrm{Km}$ patch
- one pixel corresponds to a $30\mathrm{m} \times 30\mathrm{m}$ spot

By the numbers

- two satellites, orbiting the earth one every 80mins
- every 16 days, each satellite passes over almost every square kilometer of the earth's surface
- an image covers an area approximately $185\mathrm{Km} \times 185\mathrm{Km}$
- each images contain $2,000 \times 2,000 = 4,000,000$ pixels
- 7 images per $185\mathrm{Km} \times 185\mathrm{Km}$ patch
- one pixel corresponds to a $30\mathrm{m} \times 30\mathrm{m}$ spot

Key Observation
The seven Landsat images of one fixed region typically contain much redundant information, since some features will appear in several images. Yet other features, because of their color or temperature, may reflect light that is recorded by only one or two sensors.

Principal Component Analysis
Provide a view of the data in a way that extracts information better than studying each image separately.

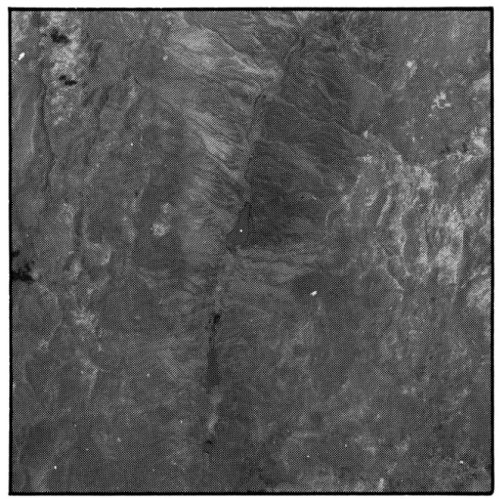
(a) Spectral band 1: Visible blue.

(b) Spectral band 4: Near infrared.

(c) Spectral band 7: Mid-infrared.

(d) Principal component 1: 93.5%.

(e) Principal component 2: 5.3%.



(f) Principal component 3: 1.2%.

Observation Matrix

data is collected in the observation matrix X

- let $X^i \in \mathbb{R}^m$ denote observation i
- each observation contains m measurements
- we collect N observations

$$X = \begin{bmatrix} X^1 & X^2 & \dots & X^N \end{bmatrix} = \begin{bmatrix} \text{---} & \text{measurement } 1 & \text{---} \\ & \vdots & \\ \text{---} & \text{measurement } m & \text{---} \end{bmatrix} \in \mathbb{R}^{m \times N}$$

Observation Matrix

data is collected in the observation matrix X

- let $X^i \in \mathbb{R}^m$ denote observation i
- each observation contains m measurements
- we collect N observations

$$X = \begin{bmatrix} X^1 & X^2 & \dots & X^N \end{bmatrix} = \begin{bmatrix} \text{---} & \text{measurement } 1 & \text{---} \\ & \vdots & \\ \text{---} & \text{measurement } m & \text{---} \end{bmatrix} \in \mathbb{R}^{m \times N}$$

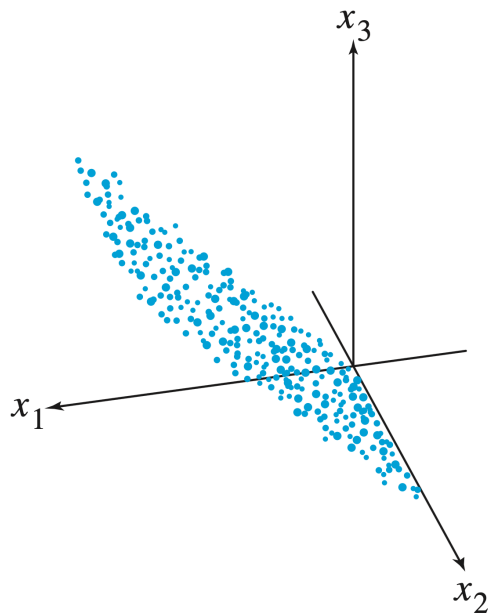
Example

we collect the age, weight, and height of students taking E4620

- 3 measurements: $m = 3$
- 25 students, observation i has the form $X^i = (\mathrm{a}_i, \mathrm{w}_i, \mathrm{h}_i)$ (variables)

$$X = \begin{bmatrix} \mathrm{a}_1 & \mathrm{a}_2 & \dots & \mathrm{a}_{25} \\ \mathrm{w}_1 & \mathrm{w}_2 & \dots & \mathrm{w}_{25} \\ \mathrm{h}_1 & \mathrm{h}_2 & \dots & \mathrm{h}_{25} \end{bmatrix}$$

Landsat observation matrix



- images (a)-(c) on p.13-7 can be viewed as a single "image" composed of 3 spectral components
- let $X^j \in \mathbb{R}^3$ denote the j^{th} pixel in the new image
- each element of X^j corresponds to pixel intensity in each of the three spectral bands
- the observation matrix has 3 rows and 4,000,000 columns

Covariance matrix

if $x = (x_1, \dots x_n)$ is a random vector with

$$\mu_i = \mathbb{E}x_i, \quad \sigma_i = \sqrt{\mathbb{E}(x_i - \mu_i)^2}, \quad \text{and} \quad \sigma_{ij} = \mathbb{E}(x_i - \mu_i)(x_j - \mu_j)$$

then

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_n^2 \end{bmatrix} = \mathbb{E}\left(\begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \\ \vdots \\ x_n - \mu_n \end{bmatrix} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \\ \vdots \\ x_n - \mu_n \end{bmatrix}^T \right)$$

is called the covariance matrix, in vector form

$$\Sigma = \mathbb{E}(x - \mu)(x - \mu)^T$$

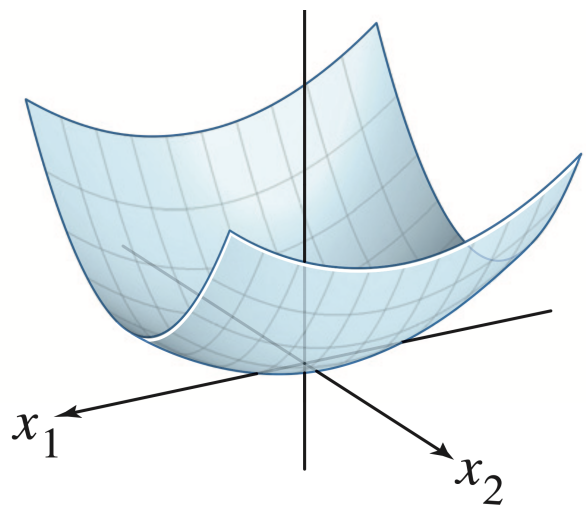
- σ_i is the standard deviation of x_i
- σ_i^2 is the variance of x_i
- σ_{ij} is the covariance of x_i and x_j and $\sigma_{ij} = \sigma_{ji}$
- $\sigma_{ij} = 0$ means x_i and x_j are not correlated

Positive semidefinite matrices

an $n \times n$ symmetric matrix, A, is associated with the quadratic form

$$q(x) = x^T Ax = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j = \sum_{i=1}^{n} a_{ii} x_i^2 + 2 \sum_{i>j} a_{ij} x_i x_j$$

- A is positive semidefinite (psd) if $q(x) \geq 0$ for all x
- A is positive definite if $q(x) > 0$ for all $x \neq 0$



- covariance matrices are always positive semidefinite

Eigenvalues of positive semidefinite matrix

we denote that A is positive semidefinite by $A \succeq 0$

$$x^T Ax \geq 0 \quad \text{for all} x \qquad \Longleftrightarrow \qquad A \succeq 0$$

- **warning:** $A \succeq 0$ does not mean (nor is it true that) $a_{ij} \geq 0$

Eigenvalues of positive semidefinite matrix

we denote that A is positive semidefinite by $A \succeq 0$

$$x^T Ax \geq 0 \quad \text{for all} x \qquad \Longleftrightarrow \qquad A \succeq 0$$

- **warning:** $A \succeq 0$ does not mean (nor is it true that) $a_{ij} \geq 0$

Theorem

The eigenvalues of a positive semidefinite matrix are all non-negative.

to see this, left multiply the eigenvalue equation by x^T:

$$Ax = \lambda x \quad \Longrightarrow \quad x^T Ax = \lambda x^T x = \lambda \|x\|_2^2 \geq 0$$

- conversely, a matrix is positive semidefinite if its eigenvalues are non-negative

PSD matrix factorization

recall from p.5-21 that if A is symmetric, then it has as eigendecomposition

$$A = Q^T \Lambda Q, \quad \text{with} \quad QQ^T = I \quad \text{and} \quad \Lambda = \mathbf{diag}(\lambda_1, \ldots, \lambda_n)$$

Theorem

If A is positive definite (semidefinite), then there exists a matrix $A^{\frac{1}{2}} \succ 0$ ($A^{\frac{1}{2}} \succeq 0$) such that $A^{\frac{1}{2}} A^{\frac{1}{2}} = A$.

PSD matrix factorization

recall from p.5-21 that if A is symmetric, then it has as eigendecomposition

$$A = Q^T \Lambda Q, \quad \text{with} \quad QQ^T = I \quad \text{and} \quad \Lambda = \mathbf{diag}(\lambda_1, \ldots, \lambda_n)$$

Theorem

If A is positive definite (semidefinite), then there exists a matrix $A^{\frac{1}{2}} \succ 0$ ($A^{\frac{1}{2}} \succeq 0$) such that $A^{\frac{1}{2}} A^{\frac{1}{2}} = A$.

Proof

As A is positive definite (semidefinite) it has an eigendecomposition

$$\begin{aligned} A &= Q^T \Lambda Q \quad \text{where} \quad QQ^T = I \\ &= Q^T \Lambda^{\frac{1}{2}} \Lambda^{\frac{1}{2}} Q \\ &= \underbrace{Q^T \Lambda^{\frac{1}{2}} Q}_{A^{\frac{1}{2}}} \underbrace{Q^T \Lambda^{\frac{1}{2}} Q}_{A^{\frac{1}{2}}}. \end{aligned}$$

□

Data pre-processing

Sample mean

the sample mean, $\hat{\mu}$, of the N observation vectors $X^1, \ldots, X^N$ is

$$\mu = \frac{1}{N}(X^1 + \cdots + X^N) = \frac{1}{N}\sum_{i=1}^{N} X^i = \frac{1}{N} X\mathbf{1}$$

$\hat{\mu}$ provides an unbiased estimate of μ, i.e., $\mathbb{E}\hat{\mu} = \mu$

Data pre-processing

Sample mean

the sample mean, $\hat{\mu}$, of the N observation vectors $X^1, \ldots, X^N$ is

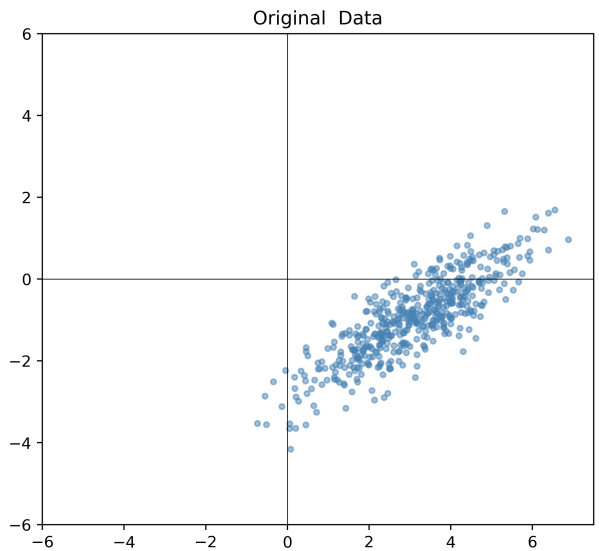
$$\mu = \frac{1}{N}(X^1 + \cdots + X^N) = \frac{1}{N}\sum_{i=1}^{N} X^i = \frac{1}{N} X\mathbf{1}$$

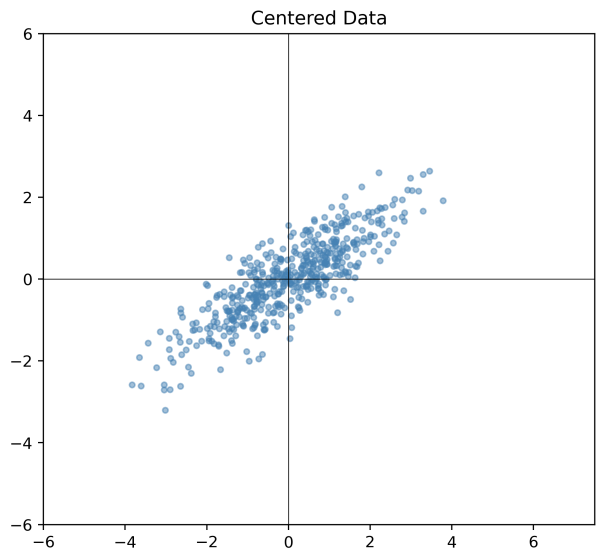
$\hat{\mu}$ provides an unbiased estimate of μ, i.e., $\mathbb{E}\hat{\mu} = \mu$

Centering

the centered observations and centered observation matrix are given by

$$\hat{X}^j = X^j - \mu, \quad j = 1, \ldots, N, \qquad \hat{X} = \begin{bmatrix} \hat{X}^1 & \hat{X}^1 & \ldots & \hat{X}^N \end{bmatrix}$$





Sample covariance

the sample covariance matrix, $S \in \mathbb{R}^{m \times m}$, is defined as

$$S = \frac{1}{N-1} \hat{X} \hat{X}^T = \frac{1}{N-1} \sum_{i=1}^{N} X^i (X^i)^T$$

note that S is the data-driven version of Σ on p.13-12

Sample covariance

the sample covariance matrix, $S \in \mathbb{R}^{m \times m}$, is defined as

$$S = \frac{1}{N-1} \hat{X} \hat{X}^T = \frac{1}{N-1} \sum_{i=1}^{N} X^i (X^i)^T$$

note that S is the data-driven version of Σ on p.13-12

Notes

the $1/(N-1)$ term is confusing, detailed discussion of where it comes from is beyond the scope of this class, but...

- $S_N = (1/N) \hat{X} \hat{X}^T$ is a more natural estimate of Σ
- S_N is a biased estimator of Σ:

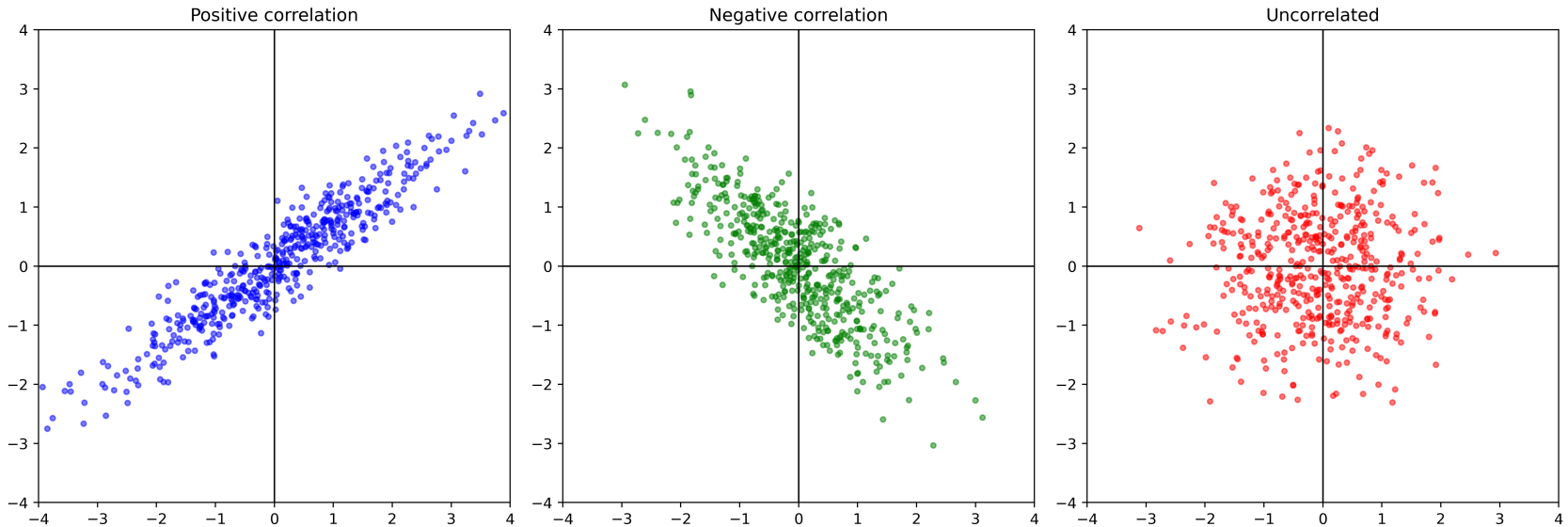
$$\mathbb{E}[S_n] \neq \Sigma, \quad \text{specifically} \quad \mathbb{E}[S_N] = \frac{N-1}{N} \Sigma$$

- S_N does not account for the fact that the sample mean is a random variable which is then used to compute S_N
- use S_N when measuring a population and use S when estimation of Σ is needed

Example

data sets generated by sampling 500 points from $\mathcal{N}(\mu, \Sigma_*)$ with, where $* \in \{+, -, \mathrm{u}\}$

$$\Sigma_+ = \begin{bmatrix} 2 & 1.5 \\ 1.5 & 1 \end{bmatrix}, \qquad \Sigma_- = \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix}, \qquad \Sigma_{\mathrm{u}} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



Sample mean and covariance

$$\hat{\Sigma}_+ = \begin{bmatrix} 1.96 & 1.35 \\ 1.35 & 1.05 \end{bmatrix}, \qquad \hat{\Sigma}_- = \begin{bmatrix} 0.96 & -0.76 \\ -0.76 & 0.94 \end{bmatrix}, \qquad \hat{\Sigma}_{\mathrm{u}} = \begin{bmatrix} 0.95 & 0.04 \\ 0.04 & 0.87 \end{bmatrix}$$

$$\hat{\mu}_+ = (0.09,\ 0.07), \qquad \hat{\mu}_- = (-0.02,\ 0.02), \qquad \hat{\mu}_{\mathrm{u}} = (-0.08,\ -0.03)$$

Covariance matrix entries

let's dig a bit deeper into the entries of a covariance matrix

- we'll work with the centered observation matrix

$$\hat{X} = \begin{bmatrix} \hat{y}_1 & \hat{y}_2 & \hat{y}_3 \\ \hat{z}_1 & \hat{z}_2 & \hat{z}_3 \end{bmatrix} = \begin{bmatrix} \hat{y}^T \\ \hat{z}^T \end{bmatrix}$$

  which corresponds to $N = 3$ observations, each containing $m = 2$ measurements

- our sample covariance matrix is

$$S_3 = \frac{1}{3}\hat{X}\hat{X}^T = \frac{1}{3}\begin{bmatrix} \|\hat{y}\|^2 & \hat{y}^T\hat{z} \\ \hat{y}^T\hat{z} & \|\hat{z}\|^2 \end{bmatrix}$$

Diagonal entries

- $S_{11} = \frac{1}{3}\|\hat{y}\|^2$ is the variance of measurement 1, expanding gives

$$S_{11} = \frac{1}{3}\left((y_1 - \hat{\mu}_1)^2 + (y_2 - \hat{\mu}_1)^2 + (y_2 - \hat{\mu}_1)^2\right) = \sigma_1^2$$

- measure of how much measurement 1 deviates from sample mean on average
- likewise, S_{22} captures variance of measurement 2

recall, we are working with the centered observation matrix

$$\hat{X} = \begin{bmatrix} \hat{y}_1 & \hat{y}_2 & \hat{y}_3 \\ \hat{z}_1 & \hat{z}_2 & \hat{z}_3 \end{bmatrix}$$

from which we constructed the biased sample covariance matrix

$$S_3 = \frac{1}{3} \begin{bmatrix} \|\hat{y}\|^2 & \hat{y}^T \hat{z} \\ \hat{y}^T \hat{z} & \|\hat{z}\|^2 \end{bmatrix}$$

Off-diagonal entries (covariances)

- $S_{12} = S_{21} = \frac{\hat{y}^T \hat{z}}{3}$, from our work on inner products p.2-16

$$\hat{y}^T \hat{z} = \|\hat{y}\| \|\hat{z}\| \cos\theta \quad \Longrightarrow \quad \cos\theta = \frac{\hat{y}^T \hat{z}}{\|\hat{y}\| \|\hat{z}\|}$$

- if $\hat{y}$ and $\hat{z}$ parallel, $\cos\theta = \pm 1$; if $\hat{y}$ and $\hat{z}$ perpendicular, $\cos\theta = 0$
- S_{12} provides a measure of how similarly $\hat{y}, \hat{z}$ deviate from their means:
  - **small** S_{12}**:** $\hat{y}, \hat{z}$ move independently
  - **positive** S_{12}**:** $\hat{y}$ and $\hat{z}$ move in the same direction
  - **negative** S_{12}**:** $\hat{y}$ and $\hat{z}$ move in the opposite directions

Principal component analysis

assume that the observation matrix $X \in \mathbb{R}^{m \times N}$ is centered

PCA

Find a new orthogonal basis for $\mathbb{R}^p$ defined by $P = [U_1\ U_2\ \ \ldots\ U_p]$, such that $y = P^{-1}x$ and the new variables $y_1, \ldots, y_p$

1. are uncorrelated
2. are arranged in decreasing order of variance

Principal component analysis

assume that the observation matrix $X \in \mathbb{R}^{m \times N}$ is centered

PCA

Find a new orthogonal basis for $\mathbb{R}^p$ defined by $P = [U_1\ U_2\ \ldots\ U_p]$, such that $y = P^{-1}x$ and the new variables $y_1, \ldots, y_p$

1. are uncorrelated
2. are arranged in decreasing order of variance

expressing the change of coordinates in matrix form:

$$\underbrace{\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}}_{X} = \underbrace{\begin{bmatrix} U_1 & U_2 & \ldots & U_p \end{bmatrix}}_{P} \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}}_{Y} \quad \Longrightarrow \quad X^k = PY^k$$

Principal component analysis

assume that the observation matrix $X \in \mathbb{R}^{m \times N}$ is centered

PCA

Find a new orthogonal basis for $\mathbb{R}^p$ defined by $P = [U_1\ U_2\ \ldots\ U_p]$, such that $y = P^{-1}x$ and the new variables $y_1, \ldots, y_p$

1. are uncorrelated
2. are arranged in decreasing order of variance

expressing the change of coordinates in matrix form:

$$\underbrace{\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}}_{X} = \underbrace{\begin{bmatrix} U_1 & U_2 & \ldots & U_p \end{bmatrix}}_{P} \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}}_{Y} \quad \Longrightarrow \quad X^k = PY^k$$

- $u_1 \ldots u_p$ form an orthonormal basis so $P^{-1} = P^T$
- $Y = P^T X \implies Y^i = U_i^T X^i$ (Y^i: measurement i in new coordinates)
- Y^i is a decomposition of X^i along the direction U_i

Constructing P

Sample covariance in new coordinates

let S_X denote the sample covariance matrix corresponding to X, then from $Y = P^T X$:

$$S_Y \;=\; \frac{1}{N-1} YY^T \;=\; \frac{1}{N-1} P^T XX^T P \;=\; P^T S_X P$$

recall, covariance matrices are symmetric, and so

$$S_X = Q\Lambda Q^T, \quad \text{with} \quad Q = [U_1 \;\; \ldots \;\; U_p], \quad \Lambda = \textbf{diag}(\lambda_1, \ldots, \lambda_p)$$

setting $P = Q$ gives $S_Y = \Lambda$

- S_Y diagonal means that covariances are zero in Y coordinates
- the orthonormal eigenvectors $U_1, \ldots, U_p$ of S_X are the loading vectors
- the principal components are $y_i = U_i^T x$, the first principal component is

$$y_1 = U_1^T x$$

Landsat example

the Landsat observation matrix had dimension $3 \times 4,000,000$ and covariance matrix:

$$S_X = \begin{bmatrix} 2382.78 & 2611.84 & 2136.20 \\ 2611.84 & 3106.47 & 2553.90 \\ 2136.20 & 2553.90 & 2650.71 \end{bmatrix}, \quad U_1 = \begin{bmatrix} .54 \\ .63 \\ .56 \end{bmatrix}, U_2 = \begin{bmatrix} -.49 \\ -.30 \\ .82 \end{bmatrix}, U_3 = \begin{bmatrix} .68 \\ -.72 \\ .14 \end{bmatrix}$$

in the new coordinates, $Y = (y_1, y_2, y_3)$, we have

$$S_Y = \begin{bmatrix} 7614.23 & & \\ & 427.63 & \\ & & 98.10 \end{bmatrix}$$

Landsat example

the Landsat observation matrix had dimension $3 \times 4,000,000$ and covariance matrix:

$$S_X = \begin{bmatrix} 2382.78 & 2611.84 & 2136.20 \\ 2611.84 & 3106.47 & 2553.90 \\ 2136.20 & 2553.90 & 2650.71 \end{bmatrix}, \quad U_1 = \begin{bmatrix} .54 \\ .63 \\ .56 \end{bmatrix}, U_2 = \begin{bmatrix} -.49 \\ -.30 \\ .82 \end{bmatrix}, U_3 = \begin{bmatrix} .68 \\ -.72 \\ .14 \end{bmatrix}$$

in the new coordinates, $Y = (y_1, y_2, y_3)$, we have

$$S_Y = \begin{bmatrix} 7614.23 & & \\ & 427.63 & \\ & & 98.10 \end{bmatrix}$$

the first principal component is

$$y_1 = 0.54x_1 + 0.63x_2 + 0.56x_3$$

Landsat example

the Landsat observation matrix had dimension $3 \times 4,000,000$ and covariance matrix:

$$S_X = \begin{bmatrix} 2382.78 & 2611.84 & 2136.20 \\ 2611.84 & 3106.47 & 2553.90 \\ 2136.20 & 2553.90 & 2650.71 \end{bmatrix}, \quad U_1 = \begin{bmatrix} .54 \\ .63 \\ .56 \end{bmatrix}, U_2 = \begin{bmatrix} -.49 \\ -.30 \\ .82 \end{bmatrix}, U_3 = \begin{bmatrix} .68 \\ -.72 \\ .14 \end{bmatrix}$$

in the new coordinates, $Y = (y_1, y_2, y_3)$, we have

$$S_Y = \begin{bmatrix} 7614.23 & & \\ & 427.63 & \\ & & 98.10 \end{bmatrix}$$

the first principal component is

$$y_1 = 0.54x_1 + 0.63x_2 + 0.56x_3$$

- the new variables are not correlated with each other
- y_1 contains more variance than y_2, which has more than y_3
- idea behind PCA is to compress by discarding variables with low variance

Total variance

PCA is most useful for compression when most of the variation in the data lies in a low dimensional subspace, i.e., can be approximately described by only a few of the variables $y_1, \ldots, y_p$

Total variance

PCA is most useful for compression when most of the variation in the data lies in a low dimensional subspace, i.e., can be approximately described by only a few of the variables $y_1, \ldots, y_p$

- let $S \in \mathbb{R}^{p \times p}$ be a covariance matrix, then the total variance is

$$\textbf{TVar}(S) = S_{11} + S_{22} + \ldots S_{pp} = \textbf{trace}(S)$$

- the variance in the observation matrix X and the transformed observation matrix Y is the same

$$\textbf{TVar}(S_X) = (S_X)_{11} + \cdots + (S_X)_{pp} = \lambda_1 + \cdots + \lambda_p = \textbf{TVar}(S_Y)$$

follows from $S_Y = P^T S_X P$ where P orthogonal

Total variance

PCA is most useful for compression when most of the variation in the data lies in a low dimensional subspace, i.e., can be approximately described by only a few of the variables $y_1, \ldots, y_p$

- let $S \in \mathbb{R}^{p \times p}$ be a covariance matrix, then the total variance is

$$\textbf{TVar}(S) = S_{11} + S_{22} + \ldots S_{pp} = \textbf{trace}(S)$$

- the variance in the observation matrix X and the transformed observation matrix Y is the same

$$\textbf{TVar}(S_X) = (S_X)_{11} + \cdots + (S_X)_{pp} = \lambda_1 + \cdots + \lambda_p = \textbf{TVar}(S_Y)$$

  follows from $S_Y = P^T S_X P$ where P orthogonal

- in Y coordinates, this motivates the quantity

$$\frac{\lambda_j}{\textbf{TVar}(Y)},$$

  the fraction of total variance captured by y_j

Landsat example

recall the Landsat covariance matrices:

$$S_X = \begin{bmatrix} 2382.78 & 2611.84 & 2136.20 \\ 2611.84 & 3106.47 & 2553.90 \\ 2136.20 & 2553.90 & 2650.71 \end{bmatrix} \quad \text{and} \quad S_Y = \begin{bmatrix} 7614.23 & & \\ & 427.63 & \\ & & 98.10 \end{bmatrix},$$

with

$$\textbf{TVar}(S_X) = \textbf{TVar}(S_Y) = \textbf{trace}(S_Y) = 8139.96$$

giving

$$\underbrace{\frac{7614.23}{8139.96} = 93.5\%,}_{\text{1st component}} \qquad \underbrace{\frac{427.63}{8139.96} = 5.3\%,}_{\text{2nd component}} \qquad \underbrace{\frac{98.10}{8139.96} = 1.2\%}_{\text{3rd component}}$$

- 93.5% of the information collected by Landsat is captured by the first principal component
- image (d) represents the data in the first principal component

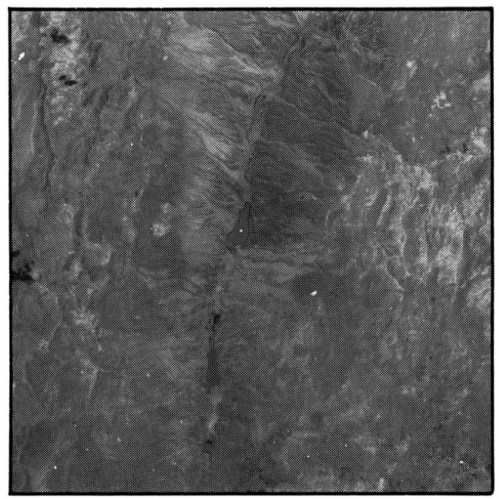
(a) Spectral band 1: Visible blue.

(b) Spectral band 4: Near infrared.

(c) Spectral band 7: Mid-infrared.

(d) Principal component 1: 93.5%.

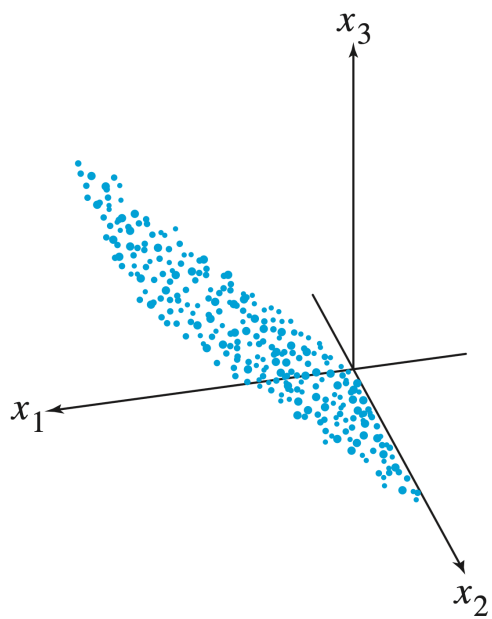
(e) Principal component 2: 5.3%.



(f) Principal component 3: 1.2%.

Geometric interpretation

- 93.5% of the information collected by Landsat is captured by the 1^{st} principal component
- the values of y_3 are all close to zero
- Geometrically, the data points lie nearly in the plane $y_3 = 0$
- y_3 can be estimated fairly accurately from y_1 and y_2
- y_2 captures very little variance either – the data is essentially 1-dimensional



Computing principal components

Summary

the eigenvectors $U_1, \ldots, U_p$ of S_X map the observed data to a new uncorrelated coordinate system:

$$\begin{bmatrix} Y^1 & Y^2 & \ldots & Y^N \end{bmatrix} = P^T \begin{bmatrix} X^1 & X^2 & \ldots & X^N \end{bmatrix}, \quad \text{with}$$

with $P = \begin{bmatrix} U_1 & U_2 & \ldots & U_p \end{bmatrix}$ and $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_p \geq 0$

the magnitude λ_i capture how much variance is captured by principal component i:

$$y_i = U_i^T x \quad \text{captures} \ \frac{\lambda_i}{\mathbf{TVar}(S_Y)}\% \ \text{of the variance}$$

- from p.13-22 we see that P can be computed via an eigendecomposition of S_X
- for large matrices, an SVD is a much more efficient and stable computation

Principal components from the SVD

Algorithm

1. collect data and store in the observation matrix $X \in \mathbb{R}^{p \times N}$
2. let $\hat{X}$ be the centered observation matrix, and define

$$W = \frac{\hat{X}^T}{\sqrt{N-1}} \qquad \Longrightarrow \qquad S_X = W^T W$$

3. compute the SVD: $W = U\Sigma V^T$, then

$$S_X = (U\Sigma V^T)^T U\Sigma V^T = V\Sigma^2 V^T$$

Notes

- the right singular vectors $v_1, \ldots v_p$ are the eigenvectors of S_X
- from our work on the SVD and eigendecomposition we know that $V\Sigma^2 V^T = Q\Lambda Q^T$

Pipeline: PCA $\rightarrow$ Clustering

1. compute principal components of X
2. keep top k components (often $k = 2$ or 3)
3. compute the PCA score matrix

$$Z = U_k^T X \in \mathbb{R}^{k \times N}$$

4. perform clustering on the columns of Z