1 INTRODUCTION

Goal of the course: Describe, understand and discover properties of data in $p > 1$ dimensions. E.g. How would we do a scatterplot in more than $p = 2$ dimensions? How to graphically represent multivariate data?

We are interested in analysing a sample of n random vectors $X_1, \ldots, X_n$, all belonging to $\mathbb{R}^p$:

$$\begin{aligned} X_1 &= (X_{11}, \ldots, X_{1p}) \in \mathbb{R}^p \\ X_2 &= (X_{21}, \ldots, X_{2p}) \in \mathbb{R}^p \\ &\vdots \\ X_n &= (X_{n1}, \ldots, X_{np}) \in \mathbb{R}^p \,. \end{aligned}$$

This means that we have a sample of n individuals, and that for each individual we observe p variables (sometimes called features).

Example:

- Consider a health study involving $n = 100$ patients.
- On each patient we measure $p = 4$ quantities: age, weight, body mass index, systolic blood pressure.
- For the ith individual, where $i = 1, \ldots, 100$ we observe

$$X_i = (X_{i1}, X_{i2}, X_{i3}, X_{i4}) .$$

- X_{i1}=age of ith patient
  X_{i2}=weight of ith patient
  X_{i3}= body mass index of ith patient
  X_{i4}=systolic blood pressure of ith patient.

- Often we gather the observations into an $n \times p$ matrix

$$\mathcal{X} = \begin{pmatrix} X_{11} & \dots & X_{1p} \\ X_{21} & \dots & X_{2p} \\ & \vdots & \\ X_{n1} & \dots & X_{np} \end{pmatrix}$$

  - ☛ Each X_{ij} is a random variable.
  - ☛ The ith row represents the p variables corresponding to the ith individual,
  - ☛ the jth column represents the jth variable for all n individuals.

- We call the value taken by $X_i = (X_{i1}, \dots, X_{ip})$ the *observed value*, or *realization*.
- Often we use a lower case to denote the observed value, i.e.
  $x_i = (x_{i1}, \dots, x_{ip})$ is the realization of the random vector X_i.

Sections 2.1, 2.2, 2.3, 2.4, 2.6 and 2.7 in Härdle and Simar.

2.1 ELEMENTARY OPERATIONS

- A matrix $A = (a_{ij}) = (a_{ij})_{1 \leq i \leq n, 1 \leq j \leq p}$ with n rows and p columns:

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1p} \\ a_{21} & a_{22} & \dots & a_{2p} \\ & \vdots & & \\ a_{n1} & a_{n2} & \dots & a_{np} \end{pmatrix}$$

- A set of k rows (or columns) $a_1, \dots, a_k$ of A are said to be linearly independent if none of them can be expressed as a nontrivial linear combination of the other $k-1$ rows (or columns), i.e.

$$\sum_{i=1}^{k} c_i a_i = 0 \Rightarrow c_1, \dots, c_k = 0 \,.$$

- Rank: The rank of a matrix A, denoted by $rank(A)$, is defined as the maximum number of linearly independent rows (or columns).

- For an n by p matrix A, we always have

$$rank(A) \leq \min(n, p) .$$

- Determinant: The determinant of a square $p \times p$ matrix A, denoted by $det(A)$ or $|A|$, is a number computed from the matrix and which plays an important role in all sorts of problems. For a 2 by 2 matrix

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

it is computed by

$$|A| = a_{11}a_{22} - a_{21}a_{12} .$$

For larger matrices, we compute this recursively. If you have forgotten how a determinant is computed, see

`https://en.wikipedia.org/wiki/Determinant`

- Inverse: If $|A| \neq 0$, the inverse of a square $p \times p$ matrix A exists. It is denoted by A^{-1} and is such that

$$AA^{-1} = A^{-1}A = I_p$$

($I_p = p \times p$ identity matrix); see Härdle and Simar, Sec 2.1, for how to compute the inverse.

- We have

$$|A^{-1}| = 1/|A| \,.$$

- Trace: The trace of a square $p \times p$ matrix A, denoted by $tr(A)$, is the sum of its diagonal elements:

$$tr(A) = \sum_{i=1}^{p} a_{ii} \,.$$

- Eigenvalues and eigenvectors of a square $p \times p$ matrix A:

  The (non zero) $p \times 1$ vector v is an eigenvector of A with eigenvalue λ if it is such that

  $$Av = \lambda v \, .$$

  Note that λ is a number (not a vector).

  - ☛ All eigenvalues satisfy

    $$|A - \lambda I_p| = 0.$$

    (they are the p roots of the above polynomial of order p in λ).

  - ☛ The eigenvalues are not necessarily all different from each other.

  - ☛ In practice we can compute them with a software, e.g. R.

  - ☛ Constant multiples of an eigenvector v with eigenvalue λ are also eigenvectors with eigenvalue λ.

- Suppose the square $p \times p$ matrix A has eigenvalues $\lambda_1, \ldots, \lambda_p$.

  Let Λ be the diagonal matrix $\Lambda = diag(\lambda_1, \ldots, \lambda_p)$ with the λ_i's on the diagonal and 0 everywhere else.

  Then we have

$$det(A) = |A| = |\Lambda| = \prod_{i=1}^{p} \lambda_i$$

  and

$$tr(A) = tr(\Lambda) = \sum_{i=1}^{p} \lambda_i.$$

- Orthogonal matrix: If A is a p by p matrix such that $AA^T = A^T A = I_p$, then A is known as an orthogonal matrix.

$\mathcal{A}(n \times n),\ \mathcal{B}(n \times n),\ c \in \mathbb{R}$

$$\operatorname{tr}(\mathcal{A} + \mathcal{B}) = \operatorname{tr}\mathcal{A} + \operatorname{tr}\mathcal{B}$$

$$\operatorname{tr}(c\mathcal{A}) = c\ \operatorname{tr}\mathcal{A}$$

$$|c\mathcal{A}| = c^n |\mathcal{A}|$$

$$|\mathcal{A}\mathcal{B}| = |\mathcal{B}\mathcal{A}| = |\mathcal{A}||\mathcal{B}|$$

$\mathcal{A}(n \times p),\ \mathcal{B}(p \times n)$

$$\operatorname{tr}(\mathcal{A}\cdot\mathcal{B}) = \operatorname{tr}(\mathcal{B}\cdot\mathcal{A})$$

$$\operatorname{rank}(\mathcal{A}) \leq \min(n, p)$$

$$\operatorname{rank}(\mathcal{A}) \geq 0$$

$$\operatorname{rank}(\mathcal{A}) = \operatorname{rank}(\mathcal{A}^\top)$$

$$\operatorname{rank}(\mathcal{A}^\top A) = \operatorname{rank}(\mathcal{A})$$

$$\operatorname{rank}(\mathcal{A} + \mathcal{B}) \leq \operatorname{rank}(\mathcal{A}) + \operatorname{rank}(\mathcal{B})$$

$$\operatorname{rank}(\mathcal{A}\mathcal{B}) \leq \min\{\operatorname{rank}(\mathcal{A}), \operatorname{rank}(\mathcal{B})\}$$

$\mathcal{A}(n \times p),\ \mathcal{B}(p \times q),\ \mathcal{C}(q \times n)$

$$\begin{aligned} \operatorname{tr}(\mathcal{ABC}) &= \operatorname{tr}(\mathcal{BCA}) \\ &= \operatorname{tr}(\mathcal{CAB}) \\ \operatorname{rank}(\mathcal{ABC}) &= \operatorname{rank}(\mathcal{B}) \quad \text{for nonsingular } \mathcal{A}, \mathcal{C} \end{aligned}$$

$\mathcal{A}(p \times p)$

$$\begin{aligned} |\mathcal{A}^{-1}| &= |\mathcal{A}|^{-1} \\ \operatorname{rank}(\mathcal{A}) &= p \quad \text{if and only if } \mathcal{A} \text{ is nonsingular.} \end{aligned}$$

Note: There is an erratum in these formulas captured from Härdle and Simar. The "subadditivity"

$$\operatorname{rank}(\mathcal{A} + \mathcal{B}) \leq \operatorname{rank}(\mathcal{A}) + \operatorname{rank}(\mathcal{B})$$

is true, but the dimensions of $\mathcal{A}$ and $\mathcal{B}$ have to match so that the addition $\mathcal{A} + \mathcal{B}$ is defined!

Spectral decomposition

- Spectral decomposition: Suppose A is a square and symmetric $p \times p$ matrix and let

  $\lambda_1, \ldots, \lambda_p$ denotes its p eigenvalues and

  $v_1, \ldots, v_p$ denote the associated $p \times 1$ eigenvectors of norm 1 and orthogonal to each other.

Note: two $p \times 1$ vectors v and w are orthogonal if

$$v^T w = \sum_{i=1}^{p} v_i w_i = 0\,.$$

Then we can always express A in the following way, which is called the spectral decomposition of A:

$$A = \sum_{j=1}^{p} \lambda_j v_j v_j^T .$$

This can also be written in matrix form, if we let

$$\Lambda = diag(\lambda_1, \ldots, \lambda_p)$$

and

$$\Gamma = (v_1 | \ldots | v_p)$$

where Γ is a $p \times p$ orthogonal matrix whose columns are the p eigenvectors:

$$A = \Gamma \Lambda \Gamma^T .$$

- In the above notation, if $A = \Gamma\Lambda\Gamma^T$ then if we take a power of A, for example A^α, we find

$$A^\alpha = \Gamma\Lambda^\alpha\Gamma^T .$$

This is because the v_j's are orthogonal and of norm 1.

For example

$$A^2 = \Gamma\Lambda\Gamma^T\Gamma\Lambda\Gamma^T = \Gamma\Lambda^2\Gamma^T .$$

This also works for negative powers if A is invertible (which happens if and only if the eigenvalues are all nonzero). For example,

$$A^{-1} = \Gamma\Lambda^{-1}\Gamma^T$$

(A^{-1}: the inverse of the matrix A).

Singular value decomposition

More generally, a similar decomposition exists for matrices that are not necessarily square matrices. In particular, any $n \times p$ matrix A with rank r can be decomposed as

$$A = \Gamma \Lambda \Delta^T ,$$

where the $n \times r$ matrix Γ and the $p \times r$ matrix Δ are column orthonormal, which means that their colomns are orthonormal, that is

$$\Gamma^T \Gamma = \Delta^T \Delta = I_r$$

and

$$\Lambda = diag(\lambda_1^{1/2}, \ldots, \lambda_r^{1/2})$$

where each $\lambda_i > 0$.

The λ_i's are the nonzero eigenvalues of the matrices AA^T or $A^T A$; the columns of Γ and Δ are, correspondingly, the r eigenvectors of these two matrices.

- A quadratic form $Q(x)$ of the p-vector $x = (x_1, \ldots, x_p)^T$ is defined by

$$Q(x) = \sum_{i=1}^{p} \sum_{j=1}^{p} a_{ij} x_i x_j = x^T A x \,,$$

where a_{ij} is the (i, j)th element of a symmetric $p \times p$ matrix A.

- If

$$Q(x) \geq 0 \text{ for all } x \neq 0$$

then the matrix A is called positive semidefinite (or non-negative definite), which is denoted by $A \geq 0$.

- However if the quadratic form satisfies

$$Q(x) > 0 \text{ for all } x \neq 0$$

then the matrix A is called positive definite, which is denoted by $A > 0$.

- $A > 0$ is equivalent to all the eigenvalues of A satisfy:

$$\lambda_1 > 0, \ldots, \lambda_p > 0\,.$$

  Then $|A| > 0$ and A^{-1} exists.

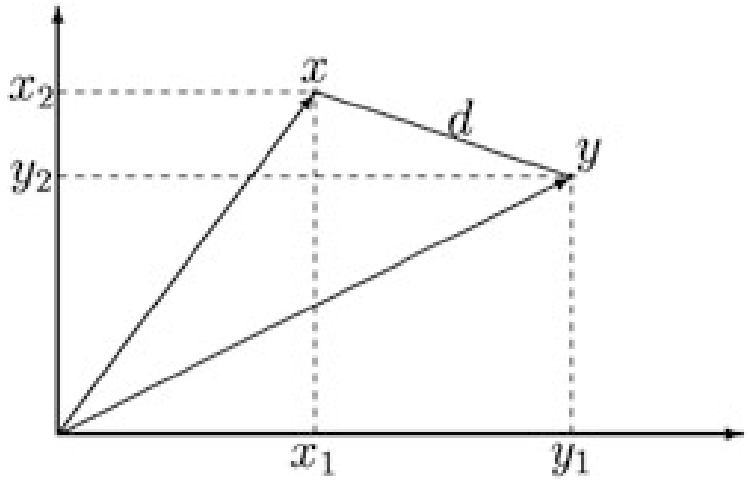
- If $A \geq 0$ and $rank(A) = r < p$, then
  - $p - r$ eigenvalues of A are equal to zero
  - while the other r are strictly positive.

Distance

- The Euclidian distance $d(x, y)$ between two vectors $x, y \in \mathbb{R}^p$ is defined by

$$d(x, y) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2} = \sqrt{(x - y)^T (x - y)}$$

Example in $\mathbb{R}^2$, where $x = (x_1, x_2)$ and $y = (y_1, y_2)$:



- A weighted version of this distance can be defined as

$$d(x,y) = \sqrt{\sum_{i=1}^{p} w_i (x_i - y_i)^2} = \sqrt{(x-y)^T W (x-y)}\,,$$

where each $w_j > 0$ and $W = diag(w_1, \ldots, w_p)$.

- This can be further generalised into the following distance:

$$d(x,y) = \sqrt{(x-y)^T A (x-y)}\,,$$

where A is a positive definite matrix.

Norm

- The (Euclidian) norm of a vector $x \in \mathbb{R}^p$ is defined by

$$\|x\| = \sqrt{\sum_{i=1}^{p} x_i^2} = \sqrt{x^T x}\,.$$

- A unit vector is a vector of norm 1.
- Multiplication by an orthogonal matrix is *norm-preserving*: If O is a p-by-p orthogonal matrix, then

$$\|Ox\| = \|x\|,$$

since $\|Ox\|^2 = x^T \underbrace{O^T O}_{\text{identity}} x = x^T x = \|x\|^2$.

- Can be generalised into a norm with respect to a positive definite matrix A:

$$\|x\|_A = \sqrt{x^T A x}\,.$$

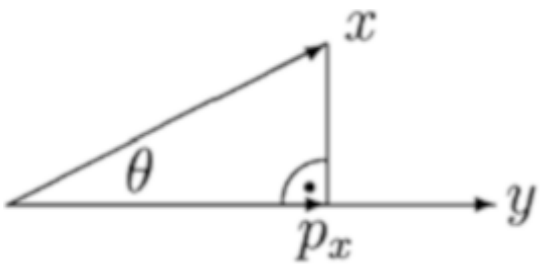
(Note: $x^T A y$ in fact defines an inner product space; see this link for instance.)

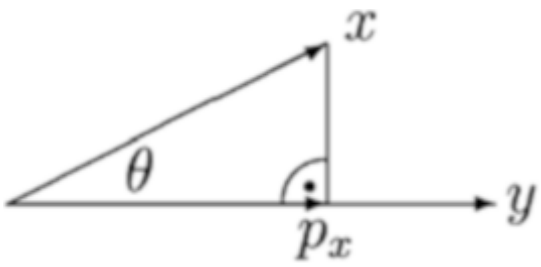
Angle between two vectors

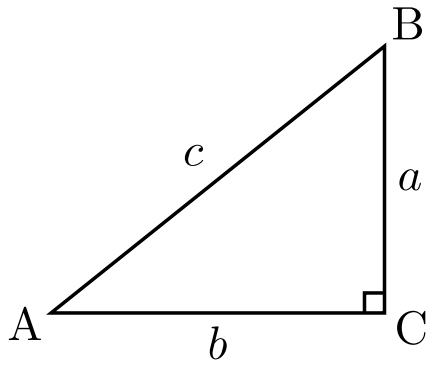
- The angle θ between two vectors $x, y \in \mathbb{R}^p$ is defined through the cosine of θ by:

$$\cos(\theta) = x^T y / \|x\| \|y\| \, .$$

Example in $\mathbb{R}^2$:







- We also know from trigonometry that, in a right angled triangle ACB with right angle at C, the cos of the angle at A is equal to length b of the segment AC divided by the length c of the segment AB. Thus

$$|\cos(\theta)| = \|p_x\|/\|x\|\,,$$

  (see figure) where p_x is called the projection of x on y.

- Since $\cos(\theta) = x^T y/(\|x\|\|y\|)$, we can find p_x by taking

$$p_x = (\cos(\theta)\|x\|)\frac{y}{\|y\|} = \frac{x^T y}{\|y\|^2}y,$$

  where $\frac{y}{\|y\|}$ is the unit vector in the direction of y.

Rotation

- When we work with vectors in $\mathbb{R}^p$, we generally describe them through a system of p axes and give the coordinates of x in that coordinate system.

- In multivariate statistics it is sometimes useful to rotate the axes (all of them at the same time) by an angle $\theta > 0$, creating in this way a new coordinate system.

- In $\mathbb{R}^2$, we can describe a rotation of angle θ via the orthogonal matrix

$$\Gamma = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix}.$$

Specifially if the original set of axes are rotated counter-clockwise through the origin by an angle θ, then the new coordinates y of a point with coordinates x in the original system of axes is given by

$$y = \Gamma x \, .$$

If the rotation of axes is clockwise, then instead we have

$$y = \Gamma^T x \, .$$

- More generally, premultiplying a vector x by an orthogonal matrix Γ with $\det(\Gamma) = 1$ geometrically corresponds to a rotation of the system of axes.

Sections 3.1, 3.2, 3.3 in Härdle and Simar.

3.1 MEAN

- The mean $\mu \in \mathbb{R}^p$ of a random vector $X = (X_1, \ldots, X_p)^T$ is defined by

$$\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix} = \begin{pmatrix} E(X_1) \\ \vdots \\ E(X_p) \end{pmatrix} .$$

- In practice we don't observe μ, but we can estimate it from a sample $X_1, \ldots, X_n$ by the sample mean

$$\bar{X} = \begin{pmatrix} \bar{X}_1 \\ \vdots \\ \bar{X}_p \end{pmatrix} ,$$

where, for $j = 1, \ldots, p$,

$$\bar{X}_j = \frac{1}{n} \sum_{i=1}^{n} X_{ij}$$

is the sample mean of the jth component X_j.

- Recall the notation

$$\mathcal{X} = \begin{pmatrix} X_{11} & \dots & X_{1p} \\ X_{21} & \dots & X_{2p} \\ & \vdots & \\ X_{n1} & \dots & X_{np} \end{pmatrix}$$

and $1_n = (1, \dots, 1)^T$, a column vector of length n.

We can express $\bar{X}$ in matrix notation as

$$\bar{X} = n^{-1} \mathcal{X}^T 1_n \, .$$

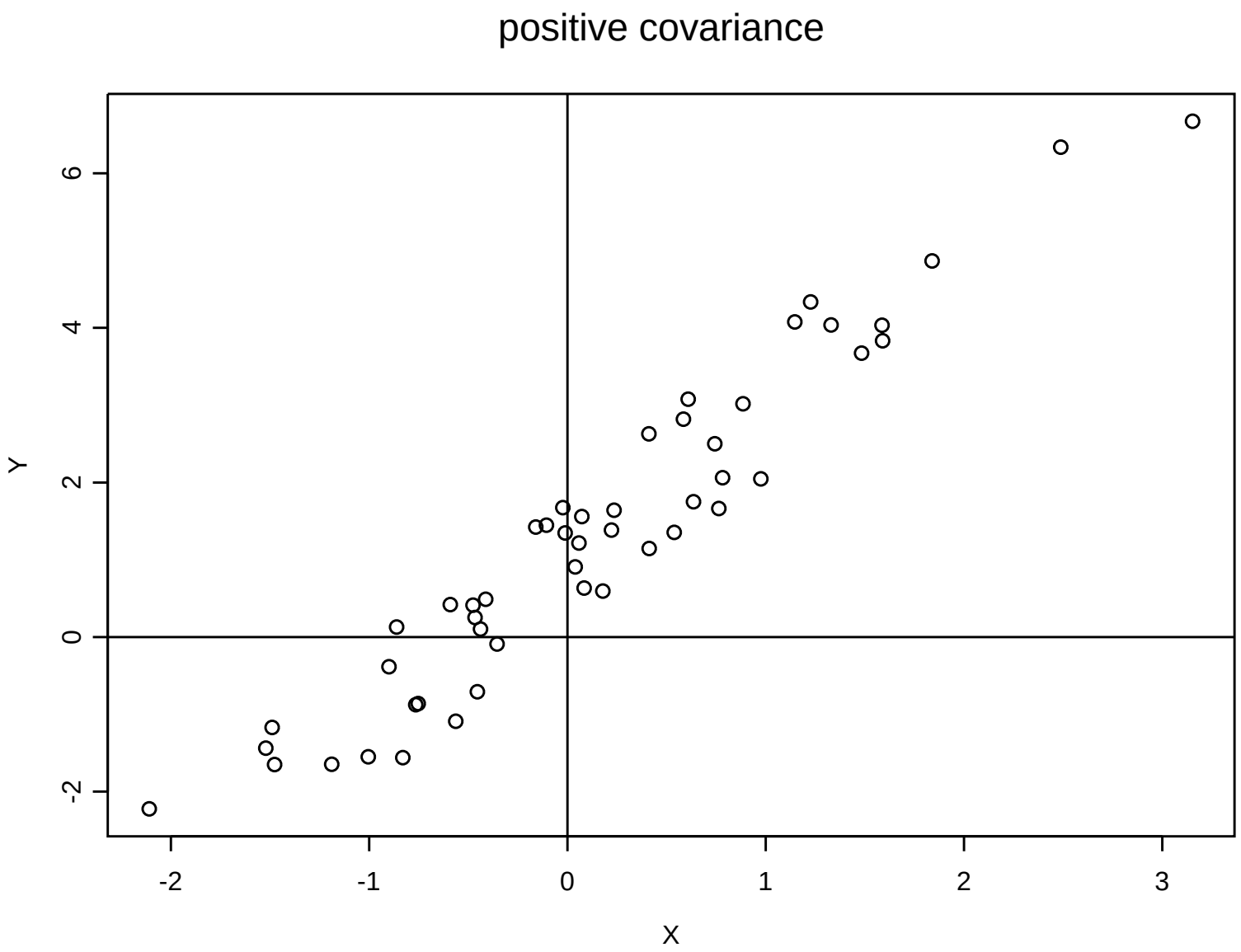
- The covariance σ_{XY} between two random variables X and Y is a measure of the linear dependence between them:

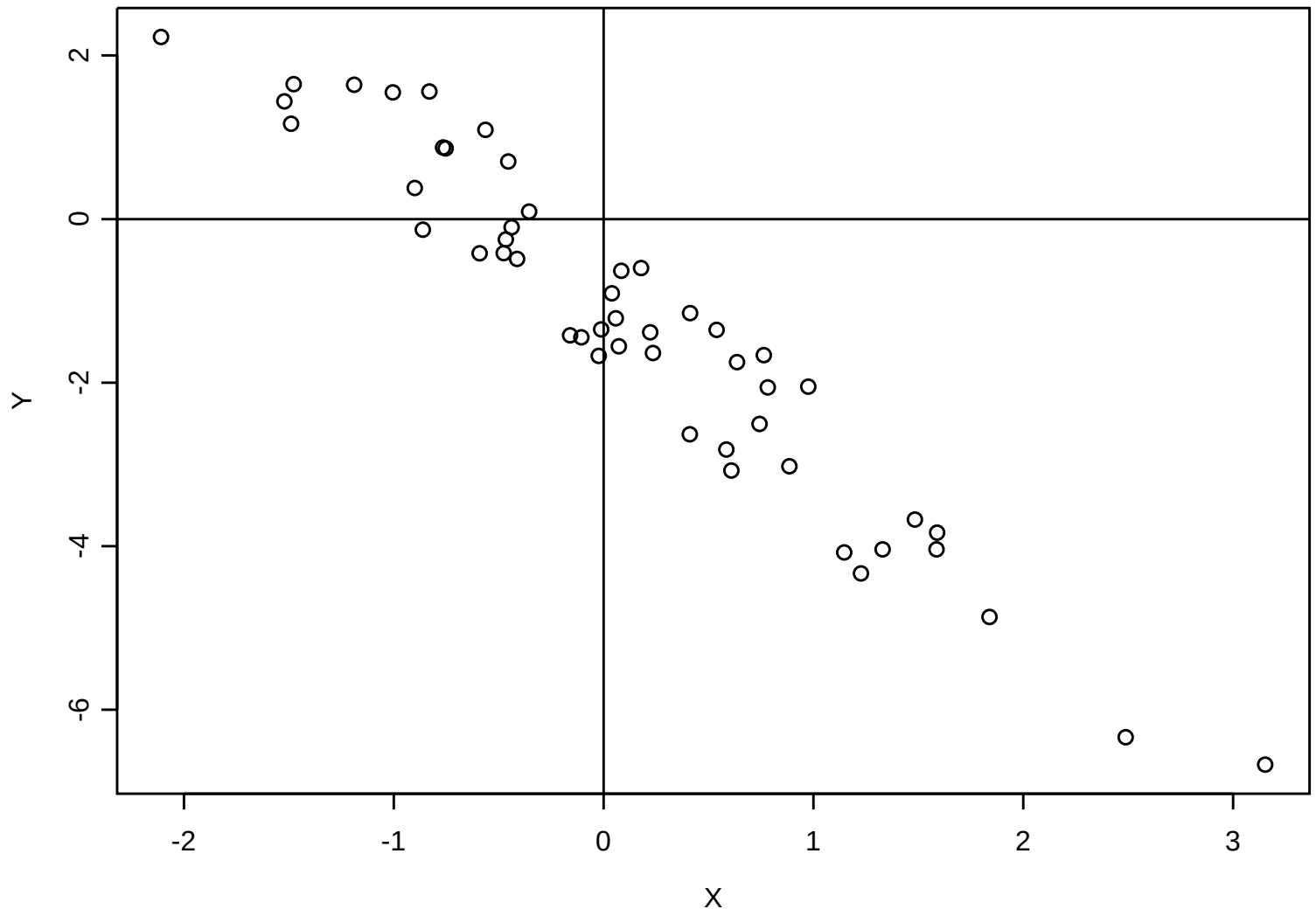
$$\sigma_{XY} = cov(X, Y) = E(XY) - E(X)E(Y)\,.$$

☛ $\sigma_{XX} = var(X)$.

☛ if X and Y are independent, then $\sigma_{XY} = 0$.

☛ However $\sigma_{XY} = 0$ does not imply that X and Y are independent (there could be nonlinear dependence).


positive covariance
Y
X
6
4
2
0
-2
-2
-1
0
1
2
3


negative covariance
Y
X
2
0
-2
-4
-6
-2
-1
0
1
2
3

If $X = (X_1, \ldots, X_p)^T$ is a p dimensional random vector, we can collect all pairwise covariances in the $p \times p$ covariance matrix Σ:

$$\Sigma = \begin{pmatrix} \sigma_{X_1X_1} & \ldots & \sigma_{X_1X_p} \\ & \vdots & \\ \sigma_{X_pX_1} & \ldots & \sigma_{X_pX_p} \end{pmatrix} = \begin{pmatrix} \sigma_{11} & \ldots & \sigma_{1p} \\ & \vdots & \\ \sigma_{p1} & \ldots & \sigma_{pp} \end{pmatrix}$$

- To highlight it is the covariance of X we can write Σ_X .
- Σ is symmetric: $\Sigma = \Sigma^T$.
- Σ is positive semi-definite: $\Sigma \geq 0$.
- In matrix notation

$$\Sigma = E\{(X - \mu)(X - \mu)^T\},$$

where X and μ are written as column p-vectors .

- In practice Σ is mostly unknown, and is estimated from the IID sample $X_1, \ldots, X_n$ by the sample covariance matrix

$$S = \begin{pmatrix} S_{X_1X_1} & \ldots & S_{X_1X_p} \\ & \vdots & \\ S_{X_pX_1} & \ldots & S_{X_pX_p} \end{pmatrix} = \begin{pmatrix} S_{11} & \ldots & S_{1p} \\ & \vdots & \\ S_{p1} & \ldots & S_{pp} \end{pmatrix},$$

where, for $j, k = 1, \ldots, p$,

$$S_{X_jX_k} = S_{kj} = \frac{1}{n-1} \sum_{i=1}^{n} (X_{ij} - \bar{X}_j)(X_{ik} - \bar{X}_k)$$

is the sample covariance between X_j and X_k.

- Again, we may write $S = S_X$ to highlight the correspondence to X.

- In matrix notation

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T = \frac{1}{n-1}\mathcal{X}^T\mathcal{X} - \frac{n}{n-1}\bar{X}\bar{X}^T$$

  where $\mathcal{X}$ is the $n \times p$ data matrix and $\bar{X}$ is the column p-vector of the sample means.

- Note that S is symmetric ($S = S^T$) and positive semidefinite

- $\mathbb{E}[S] = \Sigma$, i.e. S is unbiased for Σ. Proof:

  ☛ Without loss of generality, we can assume that $\mu = 0$ for each sample X_i, because by defining $X_{c,i} = X_i - \mu$, one can equivalently write

$$S = \frac{1}{n-1} \sum_{i=1}^{n} (X_{c,i} - \bar{X}_c)(X_{c,i} - \bar{X}_c)^T$$

  ☛ Now, one has that

$$\begin{aligned}(n-1)E[S] &= \sum_{i=1}^{n} E(X_i X_i^T) - nE(\bar{X}\bar{X}^T) \\ &= \sum_{i=1}^{n} E(X_i X_i^T) - n^{-1} E(\sum_{i=1}^{n} X_i \sum_{i=1}^{n} X_i^T) \\ &= n\Sigma - \Sigma = (n-1)\Sigma.\end{aligned}$$

- Unless otherwise specified, we will mostly consider S instead of the biased estimator

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_i)(X_i - \bar{X}_i)^T$$

- *Attention*: We have used different notation from Härdle and Simar to denote these matrices.

- Problem with covariance matrix: not unit invariant, i.e. if we change the units, covariances change.

- Correlation: a measure of linear dependence which is unit invariant.

- The correlation matrix P of a random vector $X = (X_1, \ldots, X_p)^T$ is a $p \times p$ matrix defined by:

$$P = \begin{pmatrix} 1 & \rho_{12} & \ldots & \rho_{1p} \\ \rho_{21} & 1 & \ldots & \rho_{2p} \\ & \vdots & & \\ \rho_{p1} & \rho_{p2} & \ldots & 1 \end{pmatrix}$$

where

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$$

is the correlation between X_i and X_j.

- ☛ We always have $-1 \leq \rho_{ij} \leq 1$.
- ☛ ρ_{ij} is a measure of the linear relationship between X_i and X_j.
- ☛ $|\rho_{ij}| = 1$ means perfect linear relationship.
- ☛ $\rho_{ij} = 0$ means absence of linear relationship, but does not imply independence.

- In matrix notation

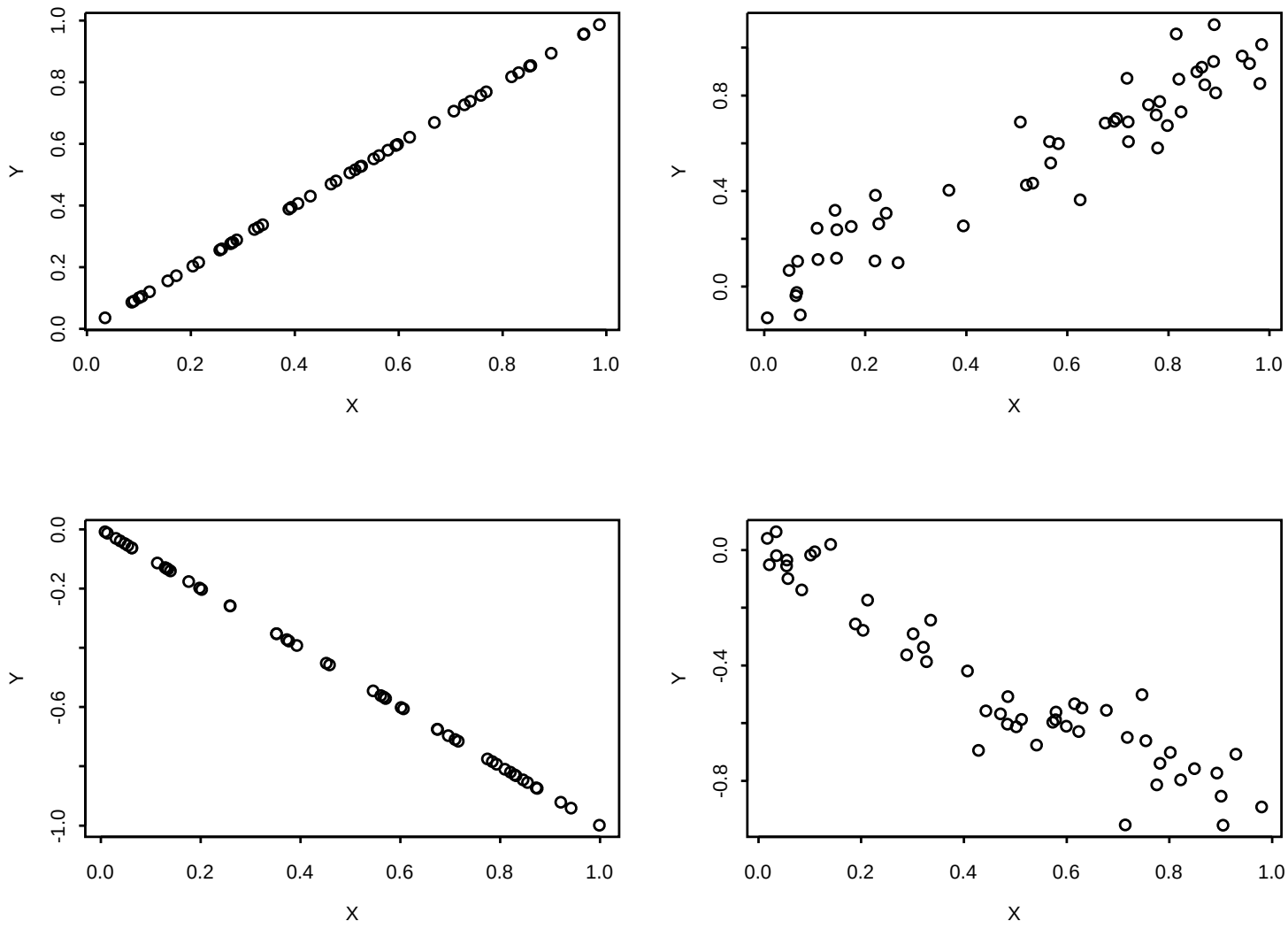
$$P = D^{-1/2} \Sigma D^{-1/2},$$

where Σ is the $p \times p$ covariance matrix and

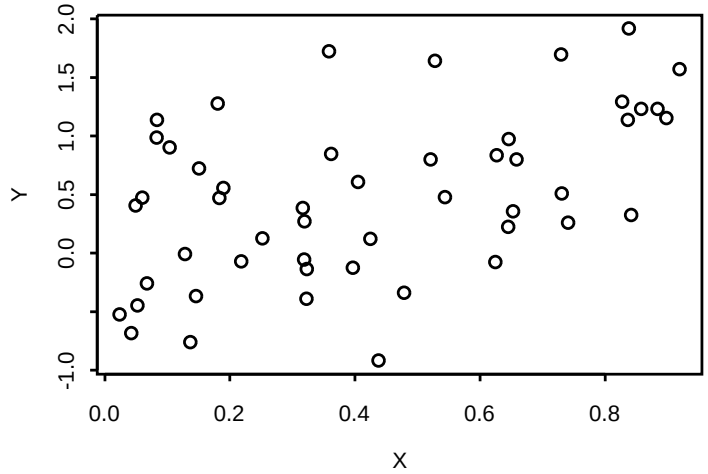
$$D = diag(\sigma_{11}, \ldots, \sigma_{pp})$$

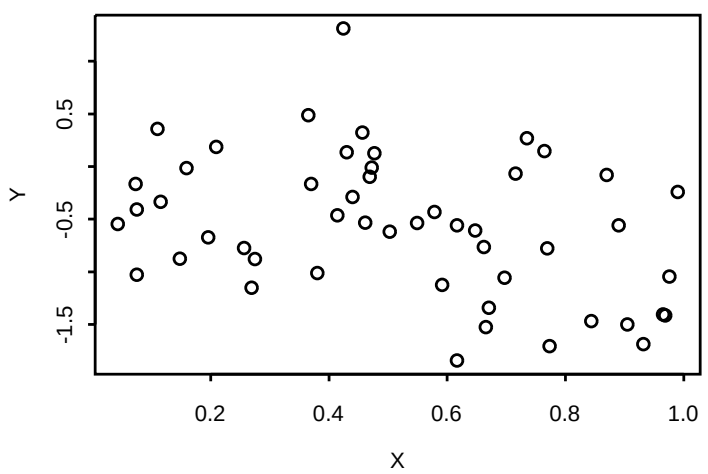
is the $p \times p$ diagonal matrix of variances.

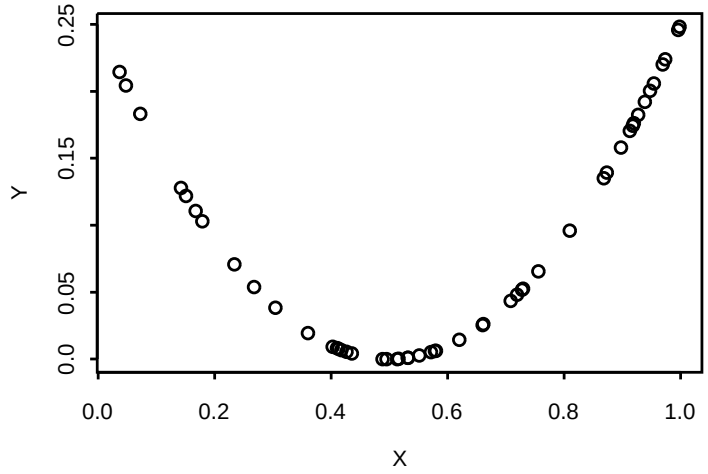
Strong positive and negative correlations:

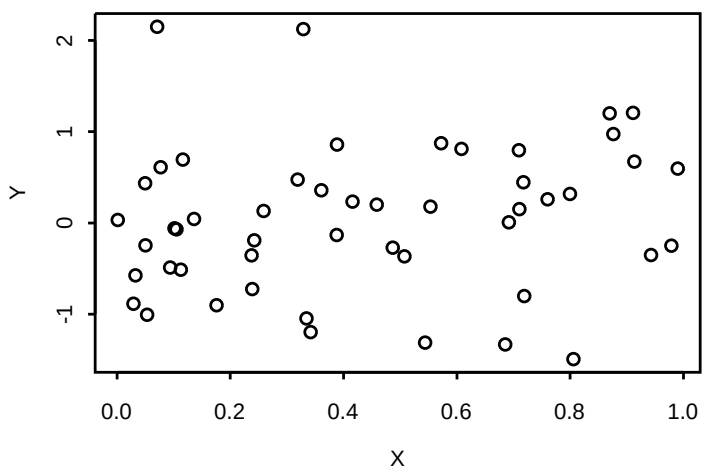


Near zero correlations:









- In practice P is mostly unknown, and is estimated from a sample $X_1, \ldots, X_n$ by the sample correlation matrix

$$R = \begin{pmatrix} R_{11} & \ldots & R_{1p} \\ & \vdots & \\ R_{p1} & \ldots & R_{pp} \end{pmatrix},$$

where, for $j, k = 1, \ldots, p$, $R_{jk} = \frac{s_{jk}}{\sqrt{s_{jj} s_{kk}}}$ is the sample correlation between X_j and X_k.

- In matrix notation we can write

$$R = D^{-1/2} S D^{-1/2},$$

where S is the $p \times p$ sample covariance matrix and, on this occasion,

$$D = diag(s_{11}, \ldots, s_{pp})$$

is the $p \times p$ diagonal matrix of sample variances.

Let $X = (X_1, \ldots, X_p)^T$ be a p-random-vector and let Y be q-random-vector defined by

$$Y = AX + b\,,$$

where A is a $q \times p$ matrix and b is a $q \times 1$ vector. Then we have

$$\begin{aligned} E(Y) &= AE(X) + b \\ \bar{Y} &= A\bar{X} + b \\ \Sigma_Y &= A\Sigma_X A^T \\ S_Y &= AS_X A^T \end{aligned}$$