p维向量

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

scalar

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

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

A is symmetric 、 X is p-vector
pxp

- If

$$Q(x) \geq 0 \;\text{ for all } x \neq (0, \dots, 0)^T$$

then the matrix A is called semi positive definite, which is denoted by $A \geq 0$. (二次型大等0 为半正定)

- However if the quadratic form satisfies

$$Q(x) > 0 \;\text{ for all } x \neq (0, \dots, 0)^T$$

二次型大于0为正定

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:

特征值全部大于0 → 正定 → 满秩，非奇异

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

特征值全大于0 正定 满秩 非奇异

Then $|A| > 0$, A^{-1} exists, A is of full rank p, A is non singular.

- If $A \geq 0$ then   不满秩 $p-r$ nonsigular, A^{-1} 不存在

$$\text{rank}(A) = r < p$$

and

  - ☛ $p - r$ eigenvalues of A are equal to zero
  - ☛ while the other r are strictly positive.

For the rest of Chap 2, vectors are columns unless specified otherwise. To get the results for rows, transpose each vector in each expression.

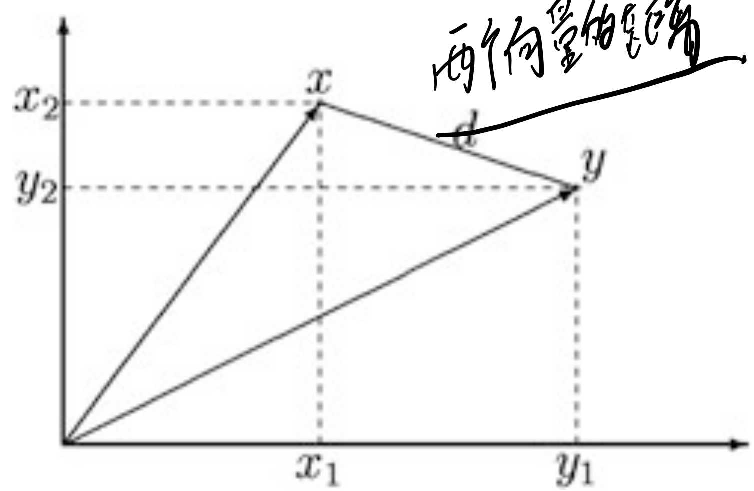
Distance

欧式距离：

- The Euclidian distance $d(x, y)$ between $x, y \in \mathbb{R}^p$ 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 $x = (x_1, x_2)^T$ and $y = (y_1, y_2)^T$:



两个向量的距离

加权

- A weighted version of this distance can be defined as

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

大小由designer决定

where each $w_i > 0$ and $W = \text{diag}(w_1, \ldots, w_p)$.

- This can be further generalised into the following distance:

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

满秩正定

where A is a $p \times p$ positive definite matrix.

Norm

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

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

(~~vec~~ vector normalization)

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

乘一个p×p的矩阵

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

模不变

since

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

$(Ox)^T(Ox) = x^T O^T O x = x^T x = \|x\|^2$

- Can be generalised into a norm with respect to a positive definite $p \times p$ matrix A:

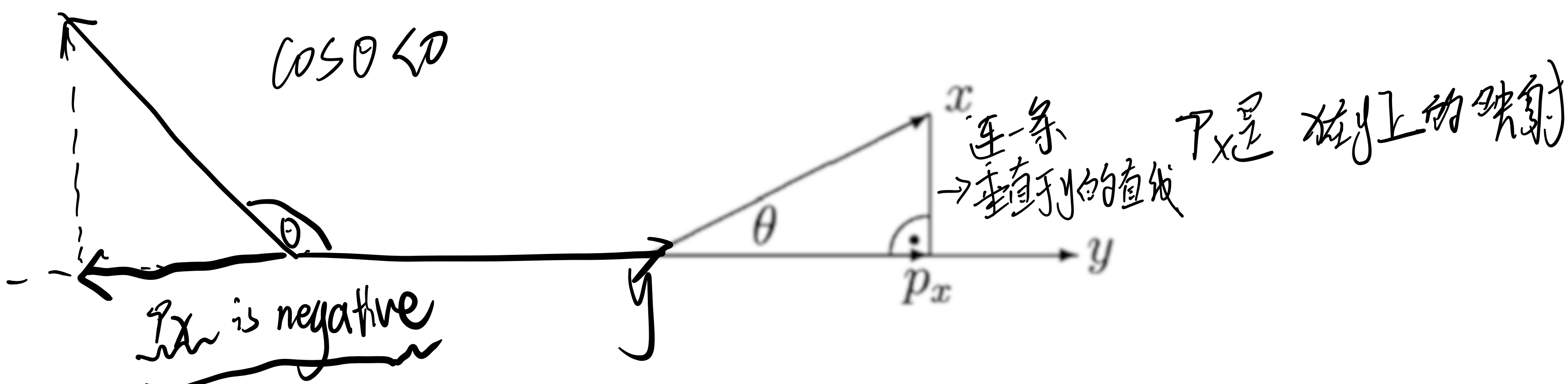
$$\|x\|_A = \sqrt{x^T A x}\,.$$

Angle between two vectors

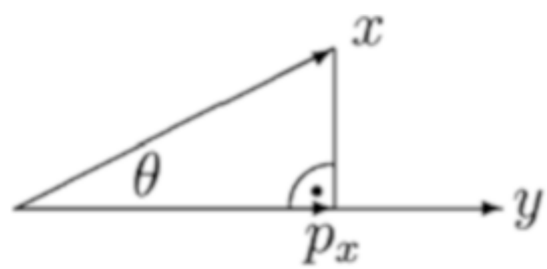
- The angle θ between two vectors $x, y \in \mathbb{R}^p$ is defined through the cosine of θ by:

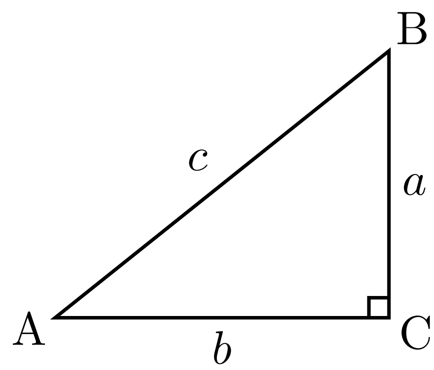
$$\cos(\theta) = \frac{x^T y}{\|x\| \cdot \|y\|}.$$

- Orthogonal projection $p_x \in \mathbb{R}^p$ of $x \in \mathbb{R}^p$ onto $y \in \mathbb{R}^p$; example in $\mathbb{R}^2$:



p_x is projected on the line defined by y. What is its length $\|p_x\|$?





- From trigono: in right angled triangle ACB with right angle at C,

$$\cos(\text{angle at A}) = b/c\,.$$

If x and y point in same direction ($x^T y > 0$):

$$\cos(\theta) = \|p_x\|/\|x\| \Rightarrow \|p_x\| = \cos(\theta)\|x\| = x^T y/\|y\|\,;$$ $\|p_x\| = x^T y/\|y\|$

if point in opposite directions ($x^T y < 0$): $\|p_x\| = -x^T y/\|y\|$.

In both cases,

→ length of p_x

$$p_x = \boxed{\frac{x^T y}{\|y\|}} \cdot \boxed{\frac{y}{\|y\|}}$$

where $y/\|y\|$ is the unit vector in the direction of y.

Rotation

- We often describe a vector in $\mathbb{R}^p$ through a system of p axes by giving the p coordinates of the vector 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 θ, creating in this way a new p coordinate system.

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

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

逆时针转θ ⇒ 得到新的向量坐标

If the original 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

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

新的 $y = \Gamma$ 乘旧 x

If the rotation is clockwise, then instead we have

顺时针

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

x前乘一个正交矩阵 → 坐标轴转动了角度.

- More generally, premultiplying a vector x by an orthogonal matrix Γ geometrically corresponds to a rotation of the system of axes.

3 MEAN, COVARIANCE, CORRELATION

Sections 3.1, 3.2, 3.3 in Härdle and Simar (2015).

3.1 MEAN

vector

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

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

- In practice don't usually know $\boldsymbol{\mu}$ but can estimate it from a sample

$$\mathbf{X}_1 = (X_{11}, \dots, X_{1p})^T, \dots, \mathbf{X}_n = (X_{n1}, \dots, X_{np})^T$$

by the sample mean

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

where, for $j = 1, \dots, p$, the sample mean

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

of the jth component X_j is an estimator of μ_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 $\mathbf{1}_n = (1, \dots, 1)^T$, a column vector of length n.

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

$$\bar{\mathbf{X}} = n^{-1} \mathcal{X}^T \mathbf{1}_n$$ 矩阵表示

- Note: in the slides, to avoid too heavy notations, when there is no ambiguity we will not use bold to denote a vector.

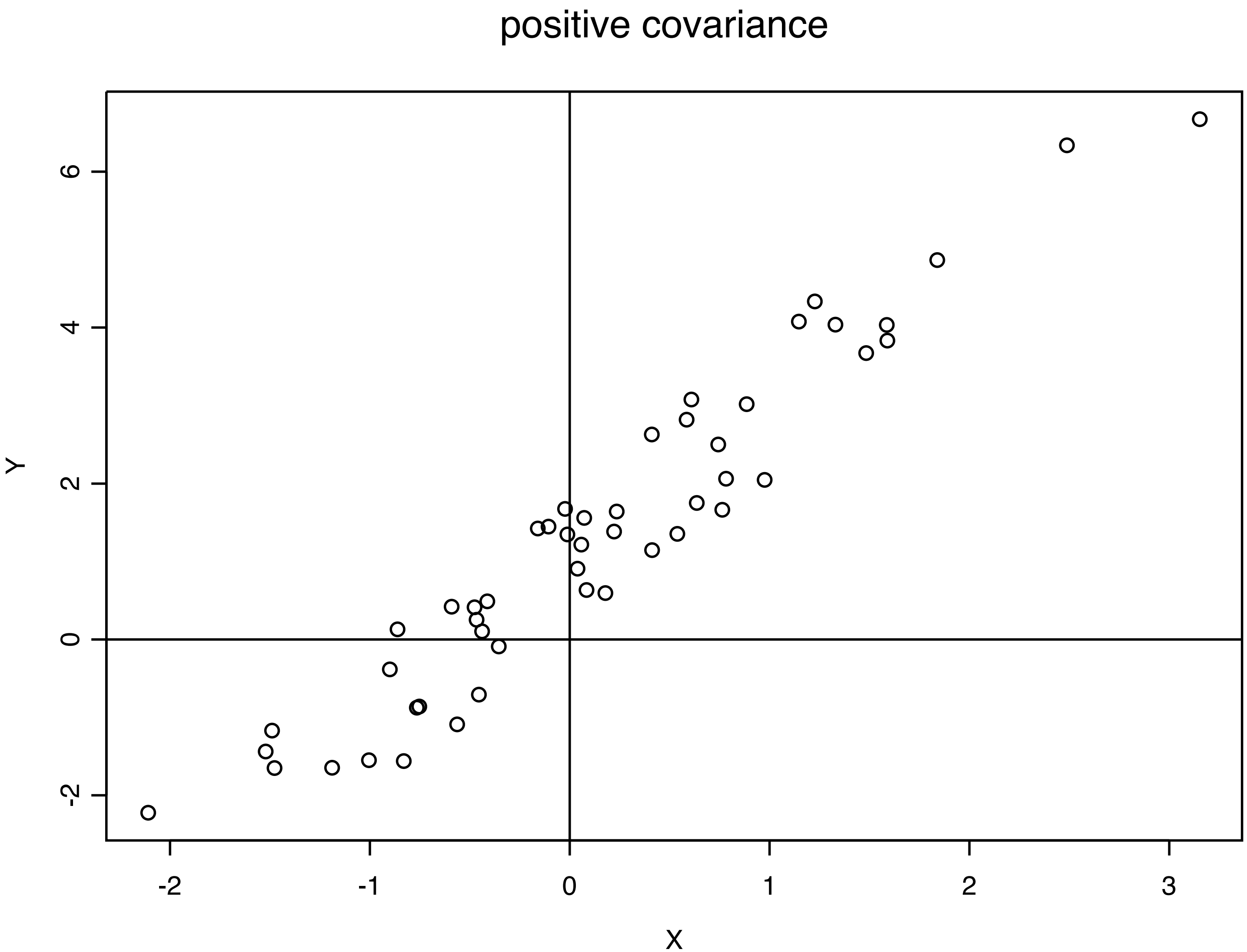
- The covariance σ_{XY} between two random variables X and Y is a measure of the linear dependence between them:

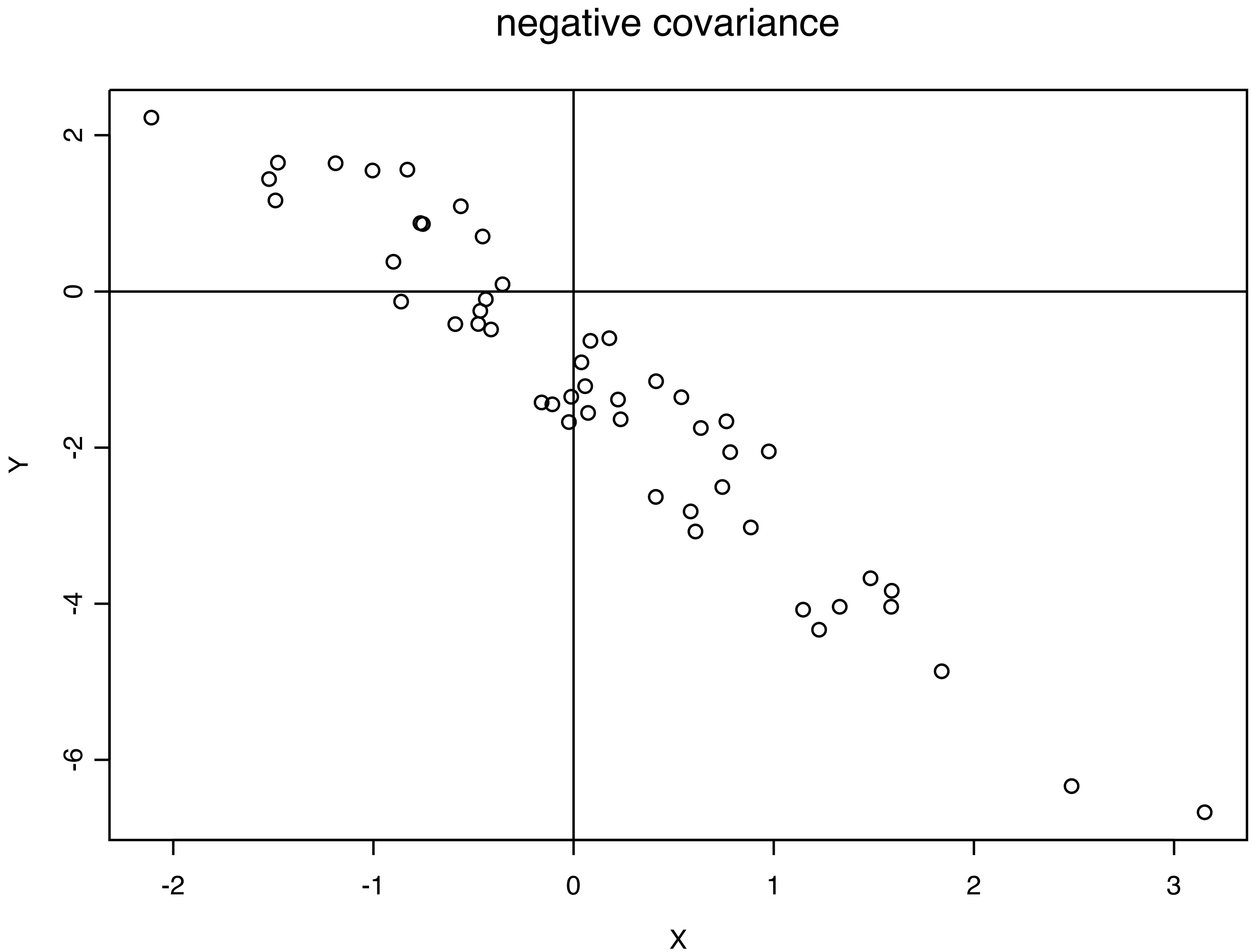
$$\sigma_{XY} = \text{cov}(X, Y) = E(XY) - E(X)E(Y)\,.$$

☛ $\sigma_{XX} = \text{var}(X)$.

☛ 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 a nonlinear dependence).


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


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

- If $\mathbf{X} = (X_1, \ldots, X_p)^T$ is a random vector, we can collect the pairwise covariances between each pair X_i and X_j 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 that it is the covariance of $\mathbf{X}$ we can write $\Sigma_{\mathbf{X}}$.

☛ Σ is symmetric: $\Sigma = \Sigma^T$.

☛ Σ is semi positive definite: $\Sigma \geq 0$.

☛ In matrix notation,

$$\Sigma = \text{var}(\mathbf{X}) = E\{(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T\} \,,$$

where $\mathbf{X}$ and $\boldsymbol{\mu}$ are written as column p-vectors .

- In practice Σ is usually unknown but can be estimated from an iid sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ 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 $j, k = 1, \ldots, p$,

unbiased

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

is the sample covariance between X_j and X_k.

- Again, we may write $S = S_{\mathbf{X}}$ to highlight the correspondence to $\mathbf{X}$.

  ☛ Like Σ, S is symmetric ($S = S^T$) and semipositive definite.

- We can obtain S by computing

$$S = \underbrace{\frac{1}{n-1}}_{p \times p} \underbrace{\mathcal{X}^T \mathcal{X}}_{p \times n \, n \times p} - \frac{n}{n-1} \underbrace{\bar{\mathbf{X}}}_{\text{empirical mean}} \bar{\mathbf{X}}^T \overset{p \times p}{,}$$

where $\mathcal{X}$ is the $n \times p$ data matrix and $\bar{\mathbf{X}}$ is written as column p-vector.

☛ Hint: always check that matrix dimensions are compatible (i.e. matrices products make sense etc).

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

单位不变

- The correlation is a measure of linear dependence which is unit invariant.

- The correlation matrix P of a random vector $\mathbf{X} = (X_1, \ldots, X_p)^T$ 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

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

ρ (covariance 的标准化)
standardized version

is the correlation between X_i and X_j.

- We always have $-1 \leq \rho_{ij} \leq 1$.

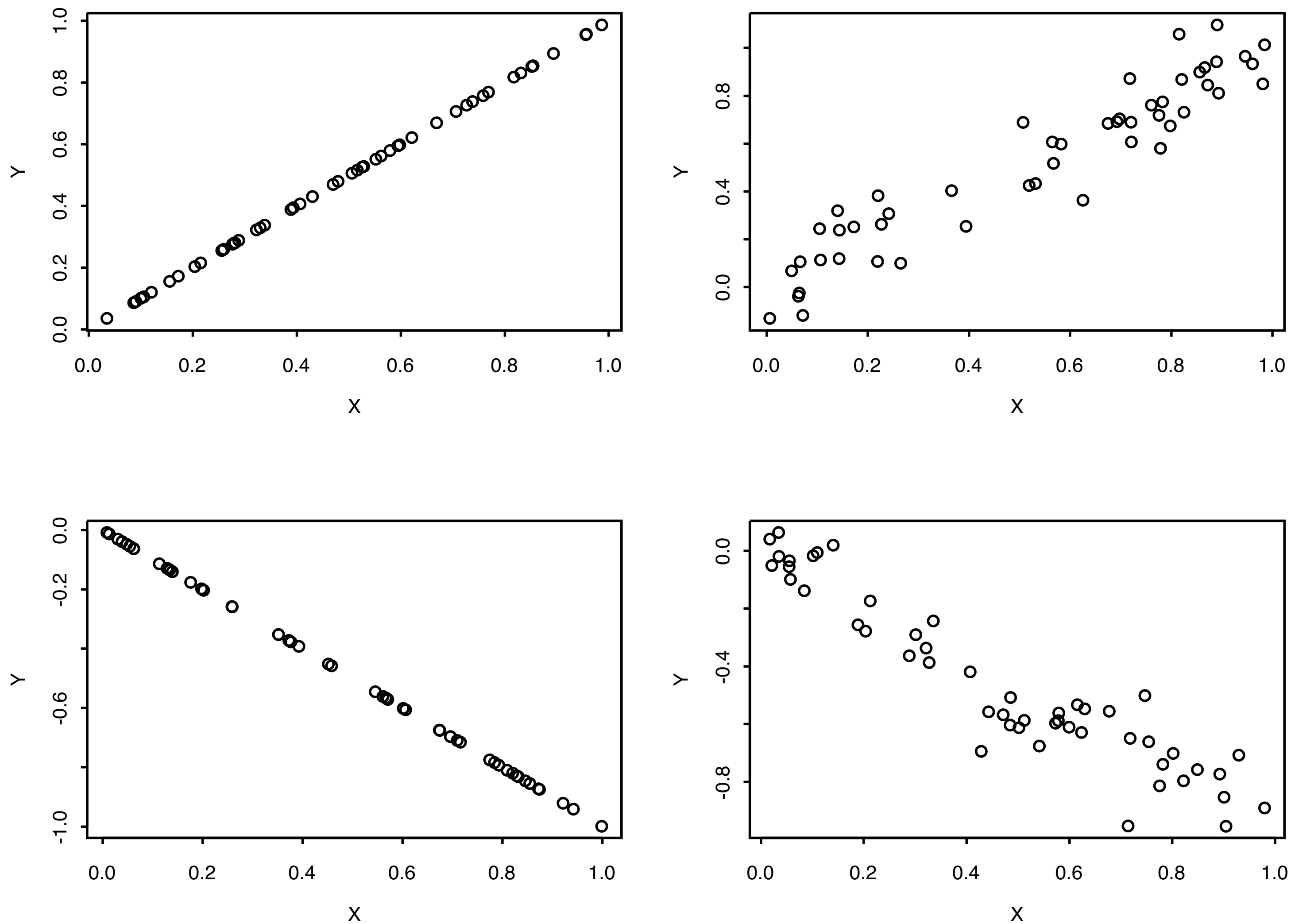
- ρ_{ij} is a measure of the linear relationship between X_i and X_j.

- $|\rho_{ij}| = 1$ means perfect linear relationship.

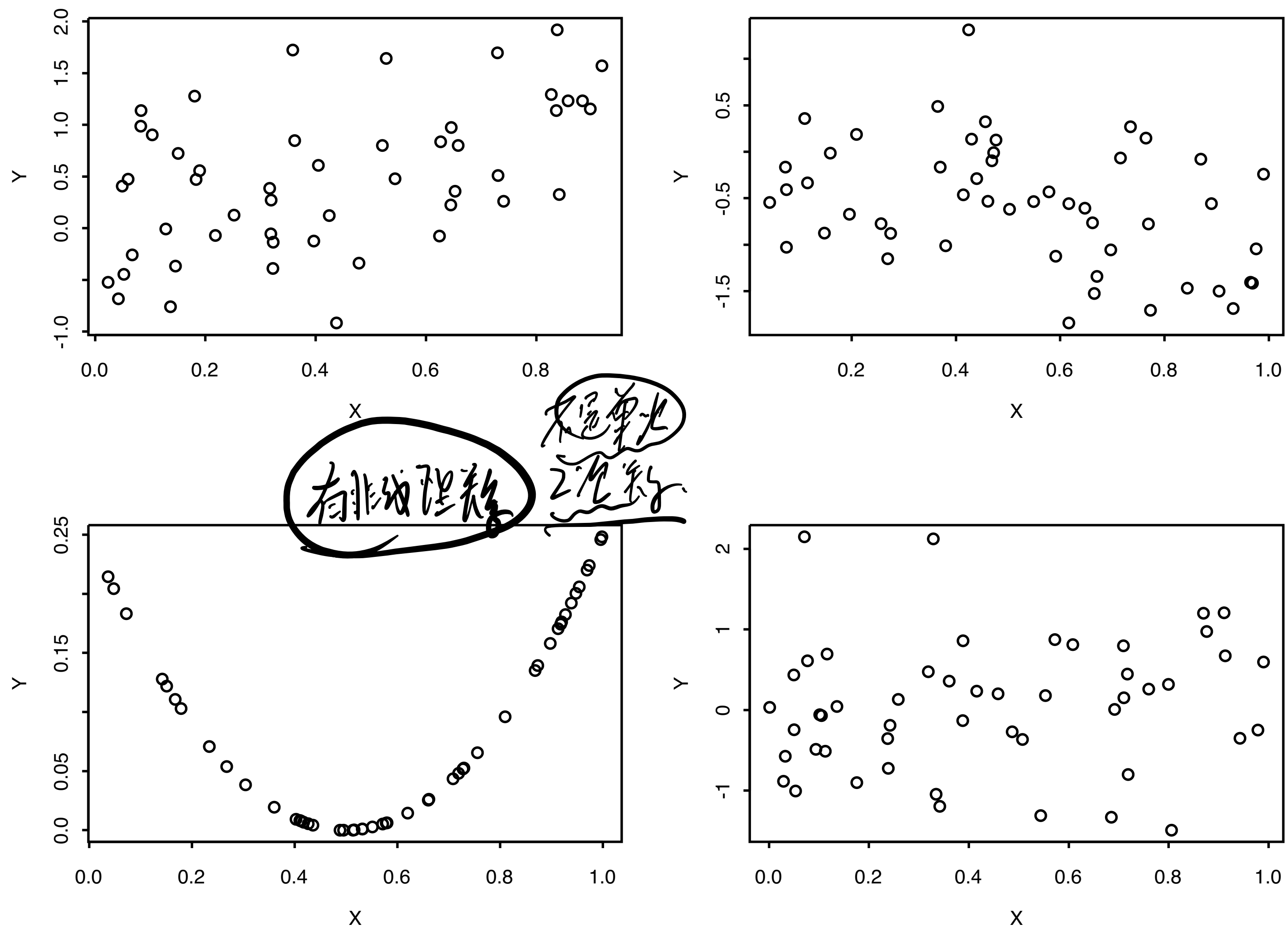
- $\rho_{ij} = 0$ means absence of linear relationship, but does not imply independence.

无线性关系 ≠ 独立

Strong positive and negative correlations:



Near zero correlations (does not always imply independence):



有非线联系

不是单独
2个系数

- We can obtain P by computing

方差
对角矩阵 开根

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

相关系数 矩阵

协方差系数矩阵

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

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

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

- In practice P is usually unknown but can be estimated from a iid sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ by the sample correlation matrix

样本的相关系数矩阵

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

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 computed from $\mathbf{X}_1, \ldots, \mathbf{X}_n$.

- In matrix notation we can write

样本相关系数矩

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

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

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

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

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

$$\mathbf{Y} = A\mathbf{X} + \mathbf{b}\,,$$

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

$$\begin{aligned}
E(\mathbf{Y}) &= A \cdot E(\mathbf{X}) + \mathbf{b} \\
\bar{\mathbf{Y}} &= A\bar{\mathbf{X}} + \mathbf{b} \\
\Sigma_{\mathbf{Y}} &= A\Sigma_{\mathbf{X}}A^T \\
S_{\mathbf{Y}} &= AS_{\mathbf{X}}A^T
\end{aligned}$$

☛ Hint: to know where to put the transpose, always check that matrix dimensions are compatible.

4 Multivariate distributions

4.1 Distribution and density function

Sections 4.1, 4.2 in Härdle and Simar (2015).

Let $\mathbf{X} = (X_1, \ldots, X_p)^T$ be a random vector.

- For all $\mathbf{x} = (x_1, \ldots, x_p)^T \in \mathbb{R}^p$, the 累积 cumulative distribution function (cdf), or distribution function, of $\mathbf{X}$ is defined by

$$F_{\mathbf{X}}(\mathbf{x}) = P(\mathbf{X} \leq \mathbf{x}) = P(X_1 \leq x_1, \ldots, X_p \leq x_p)$$

- When there is no ambiguity, we can write F instead of $F_{\mathbf{X}}$. Advantage: less heavy notations.

整的概率

- If $\mathbf{X}$ is continuous, the probability density function (pdf) or density, $f_{\mathbf{X}}$, of $\mathbf{X}$ is a nonnegative function defined through

$$F_{\mathbf{X}}(\mathbf{x}) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_p} f_{\mathbf{X}}(\mathbf{u})\, d\mathbf{u} \equiv \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_p} f_{\mathbf{X}}(u_1, \ldots, u_p)\, du_1 \ldots du_p\,,$$

where $\mathbf{u} = (u_1, \ldots, u_p)$. n维PDF 与2维类似

- It always satisfies

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{u})\, d\mathbf{u} = 1\,.$$

- When there is no ambiguity, we can write f instead of $f_{\mathbf{X}}$.

- The marginal cdf of a subset of $\mathbf{X}$ is obtained by the marginal of $\mathbf{X}$ computed at the subset, letting the other values equal to infinity.

☛ For example, the marginal cdf of X_1 is obtained by taking

marginal: X_1 X_1不动剩下的都拉满。

$$
\begin{aligned}
F_{X_1}(x_1) &= P(X_1 \le x_1) \\
&= P(X_1 \le x_1, X_2 \le \infty, \ldots, X_p \le \infty) \\
&= F_{\mathbf{X}}(x_1, \infty, \ldots, \infty)
\end{aligned}
$$

☛ and the marginal cdf of (X_1, X_3) is obtained by taking

$$
\begin{aligned}
F_{X_1,X_3}(x_1, x_3) &= P(X_1 \le x_1, X_3 \le x_3) \\
&= P(X_1 \le x_1, X_2 \le \infty, X_3 \le x_3, X_4 \le \infty \ldots, X_p \le \infty) \\
&= F_{\mathbf{X}}(x_1, \infty, x_3, \infty, \ldots, \infty)\,.
\end{aligned}
$$

CDF 求导 $\Rightarrow$ PDF

- For a continuous random vector $\mathbf{X}$, the marginal density of a subset of $\mathbf{X}$ is obtained from the joint density $f_{\mathbf{X}}$ of $\mathbf{X}$ by integrating out the other components.

  - ☛ For example, the marginal density X_1 is obtained by taking

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f_{\mathbf{X}}(x_1, u_2, \dots, u_p)\, du_2 \dots du_p$$

  - ☛ and the marginal density of (X_1, X_3) is obtained by taking

$$f_{X_1,X_3}(x_1, x_3) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f_{\mathbf{X}}(x_1, u_2, x_3, u_4, \dots, u_p)\, du_2\, du_4 \dots du_p\,.$$

- For two continuous random vectors X_1 and X_2, the conditional pdf of X_2 given X_1 is given by

$$f_{X_2|X_1}(x_2|x_1) = f_{X_1,X_2}(x_1, x_2)/f_{X_1}(x_1)\,.$$

It is defined only for values x_1 such that $f_{X_1}(x_1) > 0$.

- Two continuous random vectors X_1 and X_2 are independent if and only if

$$f_{X_1,X_2}(x_1, x_2) = f_{X_1}(x_1)f_{X_2}(x_2)\,.$$

☛ If X_1 and X_2 are independent then

$$f_{X_2|X_1}(x_2|x_1) = f_{X_1,X_2}(x_1, x_2)/f_{X_1}(x_1) = f_{X_1}(x_1)f_{X_2}(x_2)/f_{X_1}(x_1) = f_{X_2}(x_2)\,.$$

Thus knowing the value of X_1 does not change probability assessments on X_2 and vice versa.

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

$$\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix} = \begin{pmatrix} E(X_1) \\ \vdots \\ E(X_p) \end{pmatrix} = \begin{pmatrix} \int x f_{X_1}(x)\, dx \\ \vdots \\ \int x f_{X_p}(x)\, dx \end{pmatrix}.$$

☛ If X and Y are two p-vectors and α and β are constants then

$$E(\alpha X + \beta Y) = \alpha E(X) + \beta E(Y).$$

☛ If X is a $p \times 1$ vector which is independent of the $q \times 1$ vector Y then

$$E(XY^T) = E(X)E(Y^T).$$

☛ Hint: Remember to always check that matrix dimensions are compatible.

The conditional expectation $E(X_2|X_1 = x_1)$ is defined by

$$E(X_2|X_1 = x_1) = \int \underbrace{x_2 f_{X_2|X_1}(x_2|x_1)}_{} \, dx_2$$

$$f_{X_2|X_1}(x_2|x_1) = \frac{f_{X_1,X_2}(x_2,x_1)}{f_{X_1}(x_1)}$$

and the conditional covariance matrix $\text{var}(X_2|X_1 = x_1)$ is defined by

$$\text{var}(X_2|X_1 = x_1) = E(X_2 X_2^T|X_1 = x_1) - E(X_2|X_1 = x_1)E(X_2^T|X_1 = x_1),$$

if X_2 is a column vector.

☛ Hint: In doubt check the dimension of the resulting matrices to see if you get them right.

- As seen earlier, the covariance matrix Σ of a vector X of mean μ is defined by

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

We write

$$X \sim (\mu, \Sigma)$$

to denote a vector X with mean μ and covariance matrix Σ.

- We can also define a covariance matrix between a $p \times 1$ vector X of mean μ and a $q \times 1$ vector Y of mean ν by

$$\Sigma_{X,Y} = \text{cov}(X, Y) = E\{(X-\mu)(Y-\nu)^T\} = E(XY^T) - E(X)E(Y^T)\,.$$

$p \times q$

The elements of this matrix are the pairwise covariances between the components of X and those of Y.

☛ For $p \times 1$ vectors X and Y and a $q \times 1$ vector Z, we have

$$\mathrm{cov}(X + Y, Z) = \mathrm{cov}(X, Z) + \mathrm{cov}(Y, Z)$$

☛ For $p \times 1$ vectors X and Y we have

X Y 得相同是必须的条件

$$\mathrm{var}(X + Y) = \mathrm{var}(X) + \boxed{\underbrace{\mathrm{cov}(X, Y) + \mathrm{cov}(Y, X)}_{E(XY) - E(X)E(Y)}} + \mathrm{var}(Y)$$

☛ For matrices A and B and random vectors X and Y of dimensions such that the below quantities are well defined we have

$$\mathrm{cov}(AX, BY) = A\ \mathrm{cov}(X, Y)B^T\ .$$

Sections 4.4, 4.5, 5.1 in Härdle and Simar (2015).

A very useful and commonly encountered distribution is the multinormal distribution, also simply called normal distribution.

☛ Recall that in the univariate case, the density of a $N(\mu, \sigma^2)$ is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big\{ - (x-\mu)^2/(2\sigma^2)\Big\}\,.$$

☛ In the multivariate case, need to deal with vectors and matrices.

☛ The density of a normal vector $X = (X_1, \ldots, X_p)^T$ with mean $\mu = (\mu_1, \ldots, \mu_p)^T$ and positive definite covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_{11} & \ldots & \sigma_{1p} \\ & \vdots & \\ \sigma_{p1} & \ldots & \sigma_{pp} \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \ldots & \sigma_{1p} \\ & \vdots & \\ \sigma_{p1} & \ldots & \sigma_p^2 \end{pmatrix},$$

where $\sigma_j^2 = \mathrm{var}(X_j)$, is given by

$$f(x) = |2\pi\Sigma|^{-1/2} \exp\left\{ -\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu) \right\}. \tag{1}$$

$\boxed{p\text{维的列向量}}$ [ ]

☛ If the p-vector X is normal with mean μ and cov matrix Σ we write

$$X \sim N_p(\mu, \Sigma).$$

If the X_i's are independent, then

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \dots & 0 \\ & \vdots & \\ 0 & \dots & \sigma_p^2 \end{pmatrix} = \operatorname{diag}(\sigma_1^2, \dots, \sigma_p^2).$$

Thus

对角矩阵行列式的算法（对角元素连乘）

$$|2\pi\Sigma| = |\operatorname{diag}(2\pi\sigma_1^2, \dots, 2\pi\sigma_p^2)| = (2\pi)^p \sigma_1^2 \cdots \sigma_p^2$$

and

$$\Sigma^{-1} = \operatorname{diag}(\sigma_1^{-2}, \dots, \sigma_p^{-2})$$ 逆

so that

$$\begin{aligned} f(x) =& \frac{1}{\sqrt{(2\pi)^p} \prod_{j=1}^p \sigma_j} \exp\Big\{ -\frac{1}{2} \sum_{j=1}^p (x_j - \mu_j)^2 / \sigma_j^2 \Big\} \\ =& \frac{1}{\sqrt{(2\pi)^p} \prod_{j=1}^p \sigma_j} \prod_{j=1}^p \exp\Big\{ -\frac{1}{2} (x_j - \mu_j)^2 / \sigma_j^2 \Big\} \\ =& \prod_{j=1}^p \Big[\frac{1}{\sqrt{2\pi}\sigma_j} \exp\big\{ -(x_j - \mu_j)^2 / (2\sigma_j^2) \big\} \Big]. \end{aligned}$$

is the product of densities of p univariate $N(\mu_j, \sigma_j^2)$.

We see from (1) that $f(x)$ takes the same value for all $x \in \mathbb{R}^p$ such

$$(x-\mu)^T \Sigma^{-1}(x-\mu) = c$$

where c is a positive constant. For each $c > 0$, these x-values correspond to an ellipsoid (a different one for each $c > 0$; they are called contour ellipsoids).

The quantity

$$\sqrt{(x-\mu)^T \Sigma^{-1}(x-\mu)}$$

is called the Mahalanobis distance between x and μ

马氏距离（修正了欧式距离中维度不一致的问题

For example in $p = 2$ dimensions:

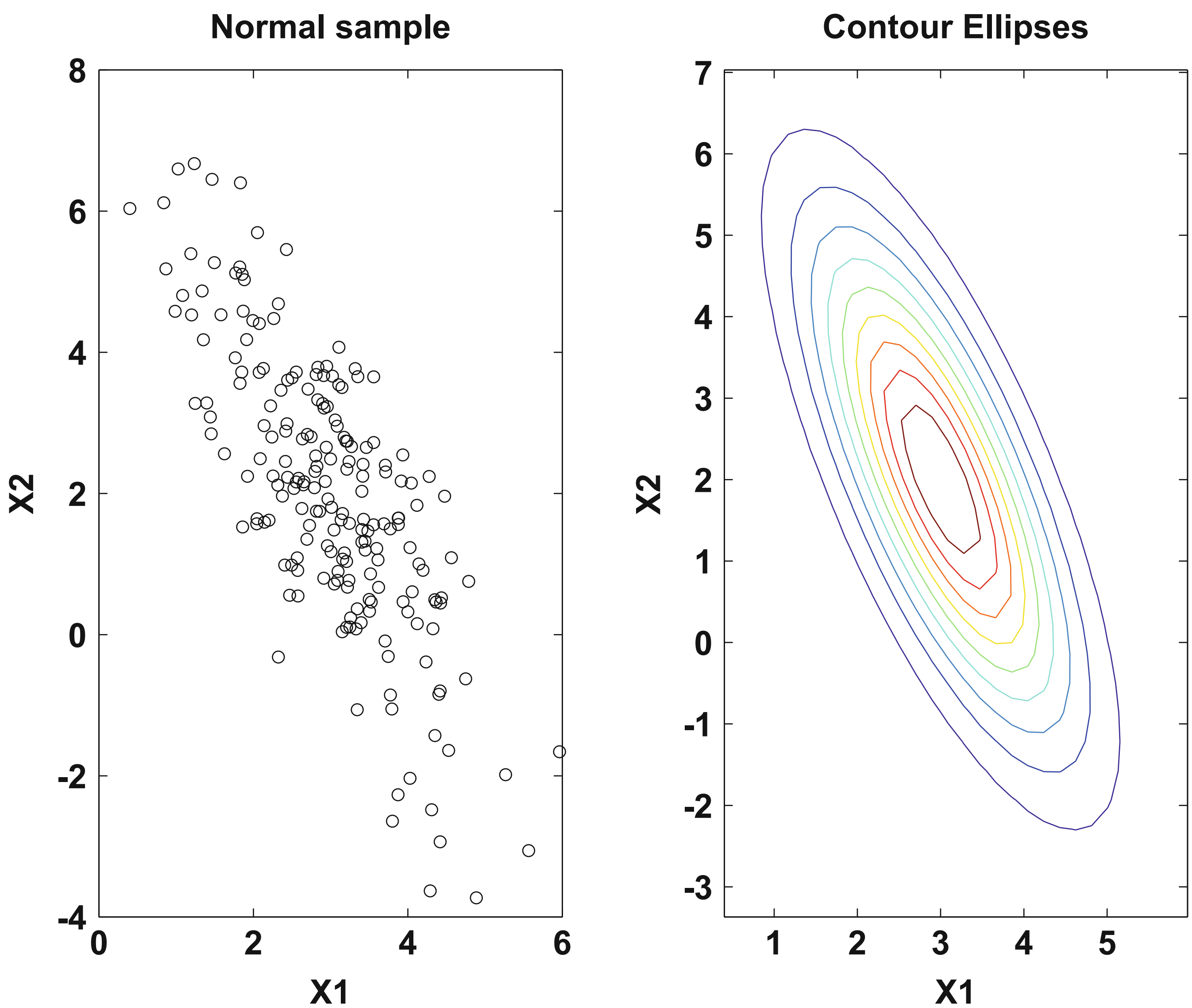


Fig. 4.3 Scatterplot of a normal sample and contour ellipses for $\mu = \binom{3}{2}$ and $\Sigma = \begin{pmatrix} 1 & -1.5 \\ -1.5 & 4 \end{pmatrix}$

☛ Let $X \sim N_p(\mu, \Sigma)$, A a $q \times p$ matrix and b a $q \times 1$ vector. Then

$$Y = AX + b \sim N_q(A\mu + b, A\Sigma A^T)\,.$$

☛ Let $X = (X_1^T, X_2^T)^T \sim N_p(\mu, \Sigma)$ where X_1 and X_2 are two column vectors. Then

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

where

$$\Sigma_{11} = \mathrm{var}(X_1), \;\; \Sigma_{22} = \mathrm{var}(X_2), \;\; \Sigma_{12} = \mathrm{cov}(X_1, X_2) \;\; \Sigma_{21} = \mathrm{cov}(X_2, X_1).$$

Then one can prove (not us)

$$\Sigma_{12} = 0 \iff X_1 \text{ and } X_2 \text{ are independent.}$$

☛ If $X \sim N_p(\mu, \Sigma)$ and A and B are matrices with p columns, then

$$AX \text{ and } BX \text{ are independent} \overset{\text{if and only if}}{\Longleftrightarrow} A\Sigma B^T = 0\,. \tag{2}$$

要记住
$A\Sigma B^T = 0 \Rightarrow AX$ 与 BX 独立

☛ If $X \sim N_p(\mu, \Sigma)$ and Σ is invertible, then

$$Y = (X-\mu)^T\Sigma^{-1}(X-\mu) \sim \chi_p^2 \tag{3}$$ (standardise the variable).

(chi square with p degrees of freedom).

$\Sigma^{-\frac{1}{2}}$

☛ If $X_1, \ldots, X_n$ are i.i.d.$\sim N_p(\mu, \Sigma)$, then

covariance matrix.

$$\bar{X} \sim N_p(\mu, \Sigma/n) \tag{4}$$

proof:

Σ Σ

$y = (x-\mu)^T \Sigma^{-\frac{1}{2}} \Sigma^{-\frac{1}{2}} (x-\mu)$

$= z^T z$

where $z = \Sigma^{-\frac{1}{2}}(x-\mu) \sim N_p(0, I_p)$

$\Sigma^{-1} = \Sigma^{-\frac{1}{2}} \Sigma^{-\frac{1}{2}}$

$\Sigma^{\frac{1}{2}}$ $\Sigma^{\frac{1}{2}}$

$\sum z_i^2 \sim \chi_p^2(1)$

 extention of χ^2

- The Wishart distribution is a generalisation to multiple dimensions of the chi square distribution.

  It depends on 3 parameters: p, a $p \times p$ scale matrix Σ and the number of degrees of freedom n:

$$W_p(\Sigma, n) \, .$$

- Recall that if $Z_1, \ldots, Z_n$ are independent $N(0,1)$ then

$$X = \sum_{k=1}^{n} Z_k^2 \sim \chi_n^2$$

  is a chi square with n degrees of freedom.

row columns 各列独立

- If M is an $p \times n$ matrix whose columns are independent and all have a $N_p(0, \Sigma)$ distribution, then the matrix

M: 各列独立

columns are normal distribution

$$\underbrace{MM^T}_{p\times p} \sim W_p(\Sigma, n) \, ,$$

行数 列数 自由度

i.e. MM^T has a Wishart distribution with parameters $\boxed{p, \Sigma \text{ and } n.}$

协方差

$W_1(\sigma^2, n)$

$$y = MM^T = (\eta_1, \cdots, \eta_n)\begin{pmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_n \end{pmatrix} = \eta_1^2 + \cdots + \eta_n^2$$

where η_j's are independent $N(0, \sigma^2)$

Thus $\eta_j = \sigma z_j$ where $z_j \sim N(0,1)$ and the z_j are independent.

Thus $M = \sigma^2 \underbrace{\sum z_j^2}_{\chi^2_n}$

143