Contours: ellipse 橢圓

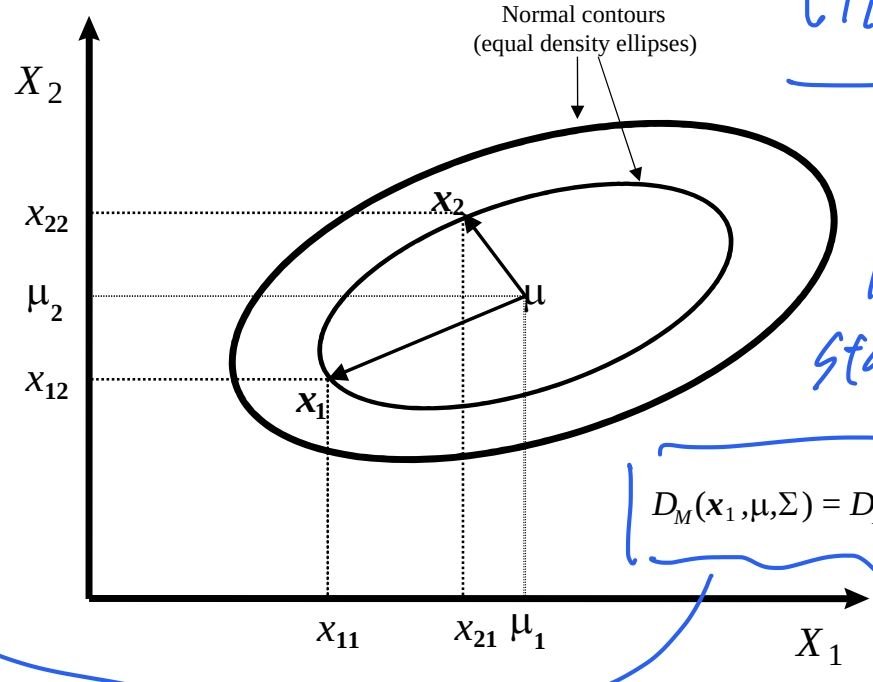
The density f of X is constant on ellipsoids of the form

$$x \in \mathbb{R}^p : \qquad (x-\mu)^T \Sigma^{-1} (x-\mu) = c.$$

The Mahalanobis distance between $x, y \in \mathbb{R}^p$

$$D_M(x, y, \Sigma) = \sqrt{(x-y)^T \Sigma^{-1} (x-y)}$$

takes into account the covariance structure of X. All points on the same ellipsoid have the same Mahalanobis distance from the center μ.



Quadratic forms

- If $Y \sim N_p(0_p, I_p)$, then

$$Y^T Y = \sum_{j=1}^{p} Y_j^2 \sim \chi_p^2,$$

where χ_p^2 denotes the chi-squared distribution with p degrees of freedom which has density

$$f(x) = \frac{1}{2^{p/2}\Gamma(p/2)} x^{p/2-1} \exp(-x/2).$$

- Consequently, for $X \sim N_p(\mu, \Sigma)$, the square of the Mahalanobis distance is distributed as a χ_p^2,

$$D^2 = (X-\mu)^T \Sigma^{-1} (X-\mu) = Y^T Y \sim \chi_p^2.$$

Conditional distribution

6. A multivariate normal variable $X \sim N_p(\mu, \Sigma)$ can be partitioned into $(X_{(1)}, X_{(2)}) \in \mathbb{R}^{p_1+p_2}$, $\mu = (\mu_{(1)}, \mu_{(2)})$ and

$$\Sigma = \begin{bmatrix} \Sigma_{(11)} & \Sigma_{(12)} \\ \Sigma_{(21)} & \Sigma_{(22)} \end{bmatrix}.$$

The conditional distribution of $X_{(2)}$ given $X_{(1)} = x_{(1)}$ can be computed explicitly by the formula

$$f_{X_{(2)}|X_{(1)}}(x_{(2)} \mid X_{(1)} = x_{(1)}) = \frac{f_{(X_{(1)}, X_{(2)})}(x_{(1)}, x_{(2)})}{f_{X_{(1)}}(x_{(1)})}.$$

It turns out that this is the density of p_2-variate normal distribution

$$N_{p_2}\left(\mu_{(2)} + \Sigma_{(21)}\Sigma_{(11)}^{-1}(x_1 - \mu_{(1)}), \Sigma_{(22)} - \Sigma_{(21)}\Sigma_{(11)}^{-1}\Sigma_{(12)}\right)$$

Central Limit Theorem

The CLT is the reason why the multivariate normal distribution is omnipresent.

- Let X_i, $i = 1, \ldots, n$ be independent and identically distributed random p-vectors with mean μ and variance Σ, then, as $n \to \infty$,

$$\sqrt{n}\left(\overline{X} - \mu\right) \xrightarrow{D} N_p(0, \Sigma)$$

where $\overline{X} = \frac{1}{n}\sum_i X_i$ and $\xrightarrow{D}$ denotes convergence in distribution.

Recall: One denotes $X_n \xrightarrow{D} X$ if $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$ for all $x \in \mathbb{R}^p$ such that F is continuous at x.

Central Limit Theorem

The CLT is the reason why the multivariate normal distribution is omnipresent.

- Let $X_i, i = 1, \ldots, n$ be independent and identically distributed random p-vectors with mean μ and variance Σ, then, as $n \to \infty$,

$$\sqrt{n}\left(\overline{X} - \mu\right) \xrightarrow{D} N_p(0, \Sigma)$$

where $\overline{X} = \frac{1}{n}\sum_i X_i$ and $\xrightarrow{D}$ denotes convergence in distribution.

Recall: One denotes $X_n \xrightarrow{D} X$ if $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$ for all $x \in \mathbb{R}^p$ such that F is continuous at x.

- This means, that for large n, $\sqrt{n}\left(\overline{X} - \mu\right)$ is approximately normal, sometimes denoted by

$$\overline{X} \dot{\sim} N_p\left(\mu, \tfrac{1}{n}\Sigma\right)$$

- The distribution of $\sqrt{n}\left(\overline{X} - \mu\right)$ is exactly normal when the X_i's are themselves normally distributed.

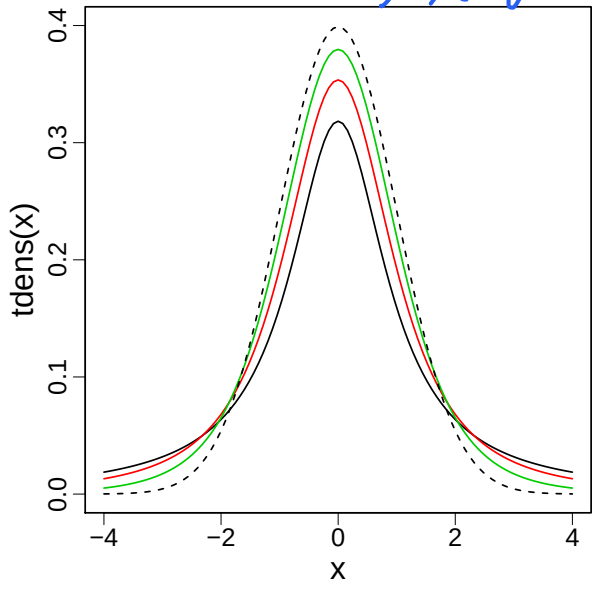
$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

Student t-distribution

- Recall the univariate Student t-distribution $X \sim t(\nu, \mu, \sigma^2)$ has density

$$f(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\sigma^2 \nu \pi}\Gamma(\frac{\nu}{2})} \left(1 + \frac{(x-\mu)^2}{\sigma^2 \nu}\right)^{-\frac{\nu+1}{2}}, \quad x \in \mathbb{R},$$

  with $E(X) = \mu$, $\nu > 1$, and $\mathrm{Var}(X) = \frac{\sigma^2 \nu}{\nu - 2}, \nu > 2$.
- The tail of the t-distribution is polynomial and thus heavier than the normal.
- When the degrees of freedom ν tends to ∞, the t-distribution converges to the normal distribution (dashed line).



多项式

$\frac{\bar{x}-\mu}{s/\sqrt{n}} \sim t(n)$

$s \leftarrow \hat{\sigma}$

一元的如何推广到多元

Multivariate Student t-distribution

- Similarly as for the multivariate normal we can apply scaling and rotation transformation to the density

$$f(x) = c_p \left(1 + \frac{(x-\mu)^T(x-\mu)}{\nu}\right)^{-\frac{\nu+p}{2}}.$$

- If X has a multivariate Student t-distribution, denoted $X \sim t_p(\nu, \mu, \Sigma)$, then its density is

$$f(x) = c_p \det(\Sigma)^{-1/2} \left(1 + \frac{(x-\mu)^T\Sigma^{-1}(x-\mu)}{\nu}\right)^{-\frac{\nu+p}{2}}$$

where ν is the degrees of freedom and

$$c_p = \frac{\Gamma((\nu+p)/2)}{(\pi\nu)^{p/2}\Gamma(\nu/2)}$$

- We have $E(X) = \mu$, $\nu > 1$, and $\operatorname{cov}(X) = \frac{\nu}{\nu-2}\Sigma, \nu > 2$.

Properties

1. The marginals of the multivariate t-distribution have univariate t-distributions, and thus heavy tails.
2. If Σ is diagonal the components of X are uncorrelated but **not** independent, unlike for the multivariate normal distribution.
3. When $\nu = 1$, it is called the multivariate Cauchy distribution.
4. Like the multivariate normal, the multivariate Student is an **elliptical distribution**: the density function only depends on the quadratic form $(x - \mu)^T \Sigma^{-1} (x - \mu)$

Wishart distribution

This is a random matrix distribution.

- Let $X_1, \ldots, X_n$ a random sample (i.i.d.) with $X_i \sim N_p(\mu, \Sigma)$, then the distribution of

$$\sum_{i=1}^{n}(X_i - \mu)(X_i - \mu)^T$$

has the Wishart distribution $W_p(n, \Sigma)$, n is the degrees of freedom.

- If $\mathbf{M} \sim W_p(n, \Sigma)$, the density is

$$f(\mathbf{M}) = \frac{|\mathbf{M}|^{(n-p-1)/2} \exp\left(-\frac{1}{2} tr(\Sigma^{-1}\mathbf{M})\right)}{2^{np/2}|\Sigma|^{n/2}\Gamma_p(n/2)}$$

where $\Gamma_p(\cdot)$ is the multivariate gamma function

$$\Gamma_p(n/2) = \pi^{p(p-1)/4} \prod_{j=1}^{p} \Gamma(n/2 + (1-j)/2)$$

and $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$ is the gamma function.

Properties, remarks and cases

- The Wishart distribution is the multivariate analogue of the chi-square distribution.
- By Cochran's theorem (cf., Mardia et al., p. 68), the scaled sample covariance is Wishart distributed

$$\sum_{i=1}^{n}(X_i - \overline{X})(X_i - \overline{X})^T = (n-1)\mathbf{S} \sim W_p(n-1, \Sigma).$$

and independent of the sample mean

$$\overline{X} = \frac{1}{n}\sum_{i=1} X_i \sim N_p(\mu, n^{-1}\Sigma).$$

- If $p = 1$, then $(n-1)\mathbf{S}$ is $\sigma^2\chi^2_{n-1}$.
- If $\mathbf{W}$ is $W_p(m, \Sigma)$, then $\mathbf{BWB}^T$ is $W_q(m, \mathbf{B}\Sigma\mathbf{B}^T)$ for any $q \times p$ matrix $\mathbf{B}$ of rank $q \leq p$.

$$\begin{cases} W \sim W_p(m, \Sigma) \\ W = \sum_{i=1}^{m} X_i^T X_i, \quad X_i \sim N_p(0, \Sigma) \end{cases}$$

$$BWB^T = B\left(\sum_{i=1}^{m} X_i^T X_i\right) B^T$$

$$BWB^T = \sum_{i=1}^{m} (BX_i^T)\cdot(BX_i^T)^T$$

it's linear

$BX_i \sim N_q(0, B\Sigma B^T)$

$q \times p \cdot p \times 1 \longrightarrow q$

by definition:

$\therefore BWB^T \sim W_q(m, B\Sigma B^T)$

Hotelling's T^2 distribution

This is a univariate distribution, an extension of the Student's t-distribution used in the context of multivariate tests.

- Let $Z \sim N(0, I_p)$ and $\mathbf{M} \sim W_p(m, I_p)$ be independent then, by definition, the distribution of

$$mZ^T\mathbf{M}^{-1}Z$$

  is Hotelling $T^2(p, m)$ (stochastic representation).

- It is equivalent to the F distribution:

$$T^2(p, m) = \{mp/(m - p + 1)\}F_{p,m-p+1}$$

Recall: The F_{d_1,d_2} distribution is the distribution of the ratio of two independent, scaled chi-squared distribution $U_1 \sim \chi^2_{d_1}$ and $U_2 \sim \chi^2_{d_2}$:

$$\frac{U_1/d_1}{U_2/d_2} \sim F_{d_1,d_2}.$$

Hotelling's T^2 distribution: properties

- If $X \sim N_p(\mu, \Sigma)$ and $\mathbf{M} \sim W_p(m, \Sigma)$ are independent, then

$$m(X-\mu)^T \mathbf{M}^{-1}(X-\mu) \sim T^2(p, m)$$

$I_p \to \Sigma$ no influence

$X_i \sim N_p(0, \Sigma)$

$M \sim W_p(m, \Sigma)$

$M^{-1} = (\Sigma^{-1/2})^T \cdot \tilde{M}^{-1} \Sigma^{-1/2}$

$M = \sum_{i=1}^{m} X_i X_i^T = \Sigma^{1/2} \left(\sum_{i=1}^{m} \tilde{X}_i \tilde{X}_i^T \right) \Sigma^{1/2 \, T}$

$\tilde{X}_i = \Sigma^{-1/2} X_i \sim N_p(0, I_p)$

$m \cdot X^T M^{-1} X = m \cdot (\Sigma^{-1/2} X)^T \cdot \tilde{M}^{-1} \cdot (\Sigma^{-1/2} X)$

by definition

$= T^2(p, m)$

I

$N_p(0, I_p)$

Hotelling's T^2 distribution: properties

- If $X \sim N_p(\mu, \Sigma)$ and $\mathbf{M} \sim W_p(m, \Sigma)$ are independent, then

$$m(X-\mu)^T \mathbf{M}^{-1}(X-\mu) \sim T^2(p, m)$$

15

- Let $X_1, \ldots, X_n$ i.i.d. $N_p(\mu, \Sigma)$, then

$$T^2 = n(\overline{X}-\mu)^T \mathbf{S}^{-1}(\overline{X}-\mu) \sim T^2(p, n-1) = \{(n-1)p/(n-p)\}F_{p,n-p},$$

Proof: Write

$$T^2 = (n-1)Z^T \mathbf{M}^{-1} Z,$$

with $Z = \sqrt{n}(\overline{X}-\mu)$ and $\mathbf{M} = (n-1)\mathbf{S}$.

$Z \sim N(0, \Sigma)$

$M \sim W_p(n-1, \Sigma)$

$T^2 = \{(n-1) \cdot [\sqrt{n}(\bar{x}-u)]^T \frac{S^{-1}}{(n-1)} \cdot \sqrt{n}(\bar{x}-u)\}$

$T^2 = n \cdot (\bar{x}-u)^T \cdot S^{-1}(\bar{x}-u)$

$\bar{X} \perp\!\!\!\perp S \Rightarrow$ independent

15

Hotelling's T^2 distribution: properties

- If $X \sim N_p(\mu, \Sigma)$ and $\mathbf{M} \sim W_p(m, \Sigma)$ are independent, then

$$m(X-\mu)^T \mathbf{M}^{-1}(X-\mu) \sim T^2(p, m)$$

- Let $X_1, \ldots, X_n$ i.i.d. $N_p(\mu, \Sigma)$, then

$$T^2 = n(\overline{X}-\mu)^T \mathbf{S}^{-1}(\overline{X}-\mu) \sim T^2(p, n-1) = \{(n-1)p/(n-p)\}F_{p,n-p},$$

Proof: Write

$$T^2 = (n-1)Z^T \mathbf{M}^{-1} Z,$$

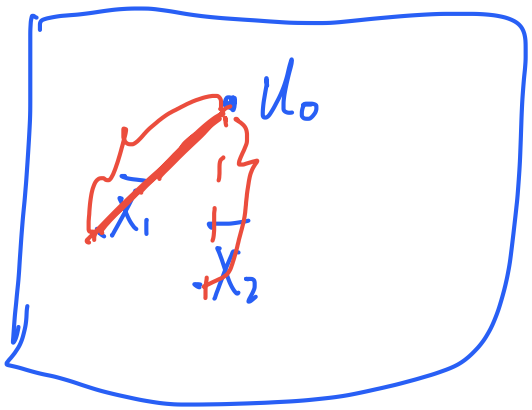
with $Z = \sqrt{n}(\overline{X}-\mu)$ and $\mathbf{M} = (n-1)\mathbf{S}$.

- This statistics (i.e. the estimated Mahalanobis distance) is a used for hypothesis testing of population means (multivariate extensions of the t-tests).

Hypothesis testing in p dimension for normal distribution. only for normal distribution.

$H_0: \mu = \mu_0 \in R^p$ [1-p 維 μ_0 is a value.] known Σ

you observe the sample mean $\bar{x}$, we want to test the true mean μ_0 if it's equal ($\bar{x}$ $\bar{x} \cong \mu_0$.



use mahalanbois distance.

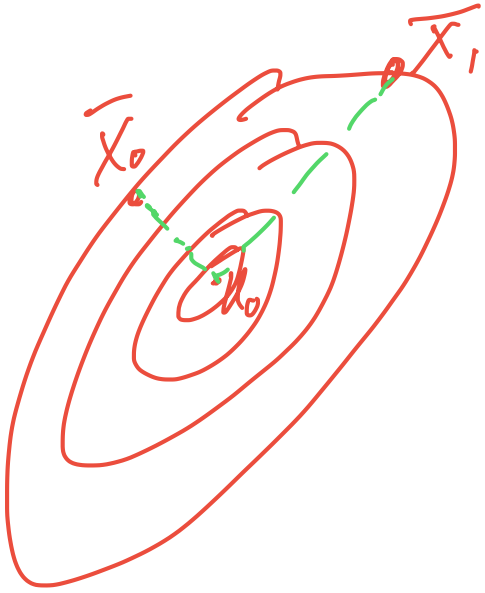
$(x-\mu_0)^T \Sigma_p^{-1} (x-\mu_0) = c$

if c is large, $\bar{x} \ncong \mu_0$

if c is small, $\bar{x} \cong \mu_0$

but we don't $\Rightarrow$ use it.

because of M-distance use the $\bar{x}$



for M-distance $\bar{x}_1$ and $\bar{x}_0$ is equal, but we think $\bar{x}_0$ is better. So we use:

$$T^2 = n \cdot (\bar{x}-\mu_0)^T \Sigma^{-1} (\bar{x}-\mu_0)$$

$$T^2 = \sqrt{n} \cdot (\bar{x}-\mu_0)^T \cdot \Sigma^{-\frac{1}{2}T} \cdot \Sigma^{-\frac{1}{2}} (\bar{x}-\mu_0)$$

$$= \underbrace{\sqrt{n} \cdot \left[\Sigma^{-\frac{1}{2}} \cdot (\bar{x}-\mu_0)\right]^T}_{\sim N(0, I_p)} \underbrace{\sqrt{n}\left[\Sigma^{-\frac{1}{2}} \cdot (\bar{x}-\mu)\right]}_{\sim N(0, I_p)} \quad \underline{\text{known } \Sigma}$$

$$= Y^T \cdot Y = \sum_{i=1}^{p} Y_i^2 \sim \chi_p^2$$

(16) $(n-1)S \sim \sigma^2 \chi_p^2$ ⇓ when $p=1$ it's a chi-squared

$\alpha = P(T^2 \geq k | H_0) = 1 - \chi_p^2(k)$ if $T^2 \geq k$, reject H_0

if unknow Σ. → positive semidefinite.

$T^2 = n\cdot(\bar{x}-\mu_0)^T S^{-1}(\bar{x}-\mu_0)$ by (17)

$\sim T^2(p, n-1)$

if $T^2 \geq k$, where $\alpha = 1 - T^2(p, n-1)(k)$ ⇒ reject k.

summary, if T^2 is large, reject.

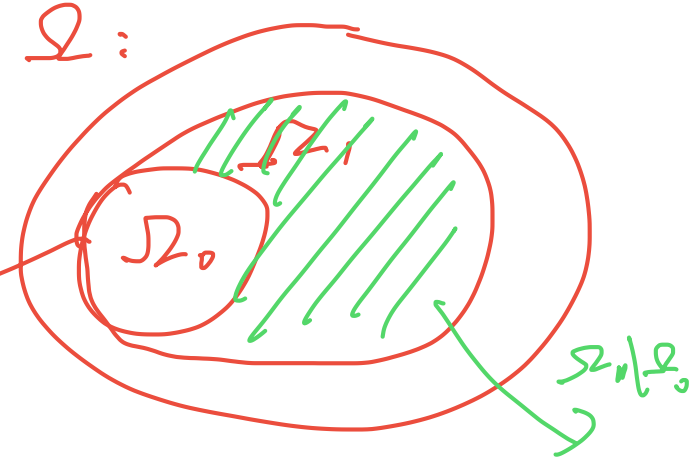
Another procedure: the likelihood ratio test – LRT

A more general framework for testing is based on the log-likelihood of the data $x_1, \ldots, x_n$ that are supposed to have density $f(\cdot; \theta)$ for some parameter $\theta \in \Omega$:

$$\ell(\theta) = \sum_{i=1}^{n} \log f(x_i; \theta) = \log L(\theta) = \log \prod_{i=1}^{n} f(x_i; \theta).$$

you can think it's covariance for example 它可以是任意变量.

we want to test if Ω_0 有 θ, 即 $\theta \in \Omega_0$ subspace



Another procedure: the likelihood ratio test – LRT

A more general framework for testing is based on the log-likelihood of the data $x_1, \ldots, x_n$ that are supposed to have density $f(\cdot; \theta)$ for some parameter $\theta \in \Omega$:

$$\ell(\theta) = \sum_{i=1}^{n} \log f(x_i; \theta) = \log L(\theta) = \log \prod_{i=1}^{n} f(x_i; \theta).$$

Likelihood ratio test: Suppose we want to test

$$H_0 : \theta \in \Omega_0, \text{ versus } H_1 : \theta \in \Omega_1 \setminus \Omega_0,$$

where $\Omega_0 \subset \Omega_1 \subset \Omega$, $dim(\Omega_0) = r$ and $dim(\Omega_1) = q$. Define the (log)-likelihood ratio

$$LRT = -2 \log(\widehat{L}_0 / \widehat{L}_1) = 2(\widehat{\ell}_1 - \widehat{\ell}_0),$$

where $\widehat{\ell}_0 = \max_{\theta \in \Omega_0} \ell(\theta)$ and $\widehat{\ell}_1 = \max_{\theta \in \Omega_1} \ell(\theta)$. Under mild conditions, for every true $\theta \in \Omega_0$:

$$LRT \overset{n \to \infty}{\longrightarrow}_D \chi^2_{q-r},$$

sometimes written as $LRT \dot{\sim} \chi^2_{q-r}$.

Another procedure: the likelihood ratio test – LRT

A more general framework for testing is based on the log-likelihood of the data $x_1, \ldots, x_n$ that are supposed to have density $f(\cdot; \theta)$ for some parameter $\theta \in \Omega$:

$$\ell(\theta) = \sum_{i=1}^{n} \log f(x_i; \theta) = \log L(\theta) = \log \prod_{i=1}^{n} f(x_i; \theta).$$

Likelihood ratio test: Suppose we want to test

$$H_0 : \theta \in \Omega_0, \text{ versus } H_1 : \theta \in \Omega_1 \setminus \Omega_0,$$

where $\Omega_0 \subset \Omega_1 \subset \Omega$, $dim(\Omega_0) = r$ and $dim(\Omega_1) = q$. Define the (log)-likelihood ratio

$$LRT = -2 \log(\widehat{L}_0 / \widehat{L}_1) = 2(\widehat{\ell}_1 - \widehat{\ell}_0),$$

where $\widehat{\ell}_0 = \max_{\theta \in \Omega_0} \ell(\theta)$ and $\widehat{\ell}_1 = \max_{\theta \in \Omega_1} \ell(\theta)$. Under mild conditions, for every true $\theta \in \Omega_0$:

$$LRT \overset{n \to \infty}{\longrightarrow}_D \chi^2_{q-r},$$

sometimes written as $LRT \dot{\sim} \chi^2_{q-r}$.

- This provides an asymptotically motivated test that does not require normally distributed samples.

example:

$X_1 \cdots X_n \sim Exp(\lambda)$

test : $H_0: \lambda = \lambda_0$

$\Omega_0 = \{\lambda_0\}$ $\Omega_1 = R$.

$q = 1$ $r = 0$

$LRT \sim \chi^2_n$

2//

$f(x;\lambda) = \lambda \cdot \exp(-\lambda x)$, $x > 0$

$\log f(x;\lambda) = \log(\lambda) - \lambda x_i$

$l(\lambda) = \sum_{i=1}^{n} \log(f(x_i;\lambda)) = n \cdot \log\lambda - n \cdot \bar{x} \cdot \lambda$

Ω_0: $\hat{l}(\lambda_0) = n \cdot \log \lambda_0 - n \cdot \bar{x} \cdot \lambda_0 \Leftarrow \max_{\lambda \in \Omega_0} l(\lambda)$

Ω_1 $\frac{\partial l(\lambda)}{\partial \lambda} = 0 \Rightarrow \frac{n}{\lambda} - n \cdot \bar{x} = 0 \Leftarrow \max_{\lambda \in \Omega_1} l(\lambda)$

$\lambda = \frac{1}{\bar{x}}$

$\hat{l}(\lambda_1) = n \cdot \log\left(\frac{1}{\bar{x}}\right) - n \cdot \frac{1}{\bar{x}} \cdot \bar{x}$

$= -n \log \bar{x} - n$

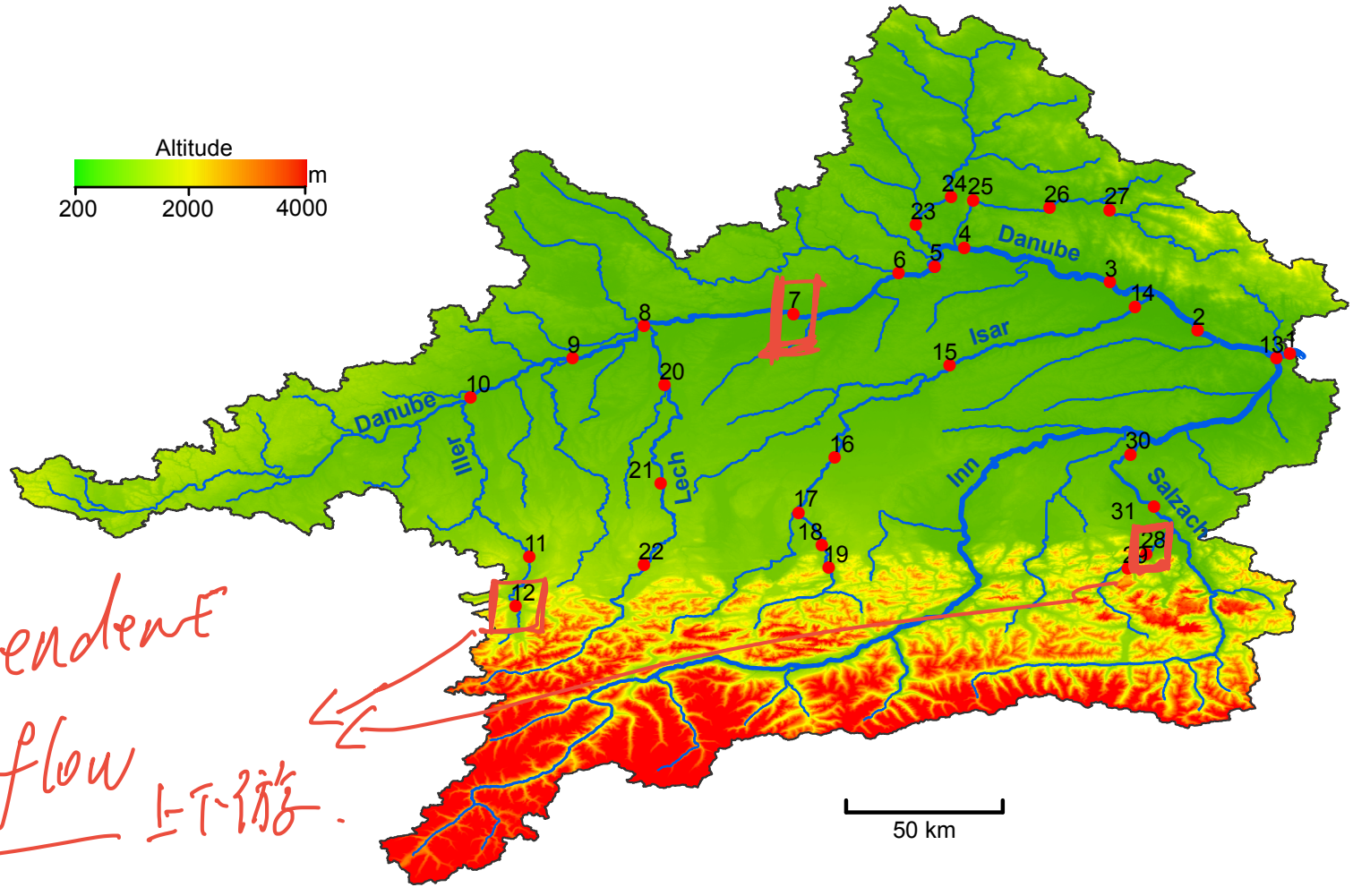
$LRT = 2\left(\hat{l}(\lambda_0) - \hat{l}(\lambda_1)\right) = 2n \cdot \left[\lambda_0 \bar{x} - 1 - \log\left(\frac{\lambda_0}{\bar{x}}\right)\right]$

n is small // close to 0, reject H_0

$\approx \chi^2_1$, for $n \to \infty$

River discharge data

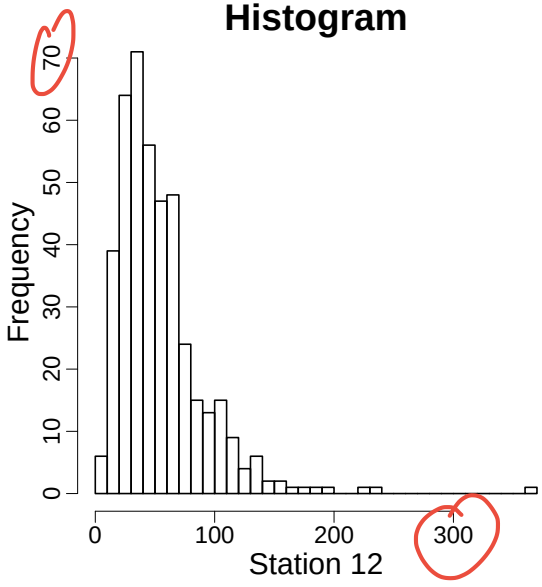
排放

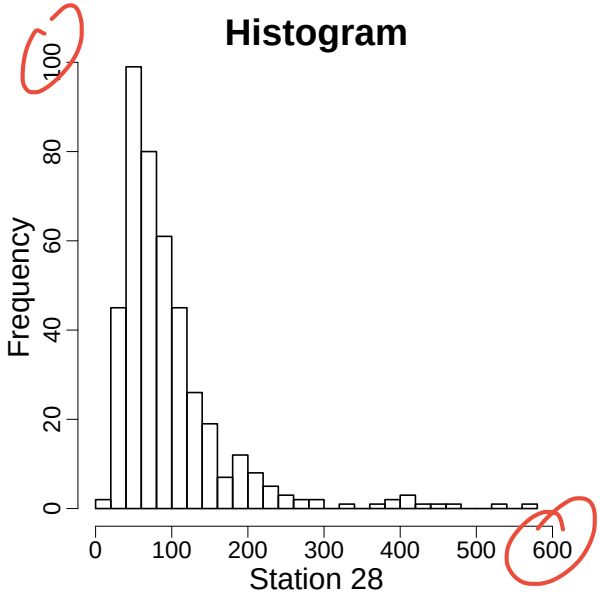


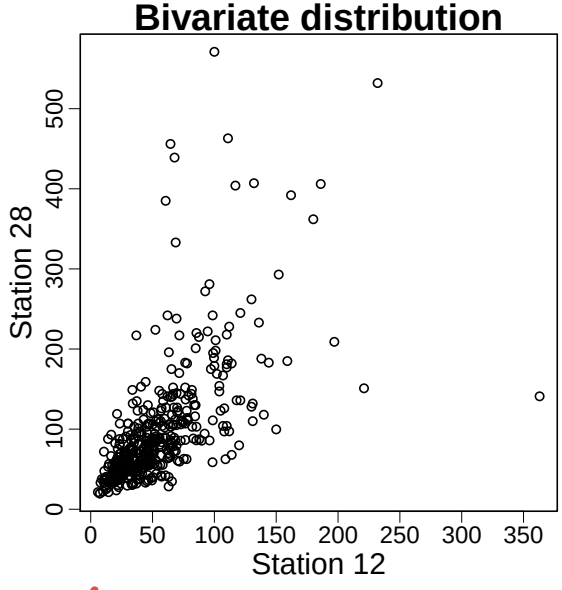
① dependent waterflow 上下游.

rain 下雨覆盖范围包含3(12, 28)

River discharge data



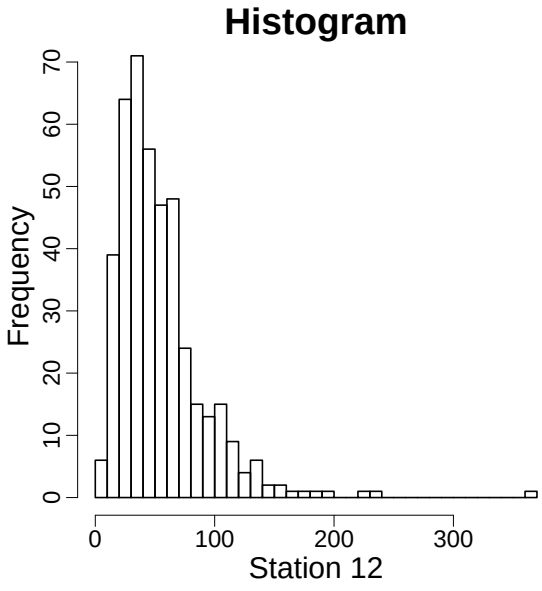


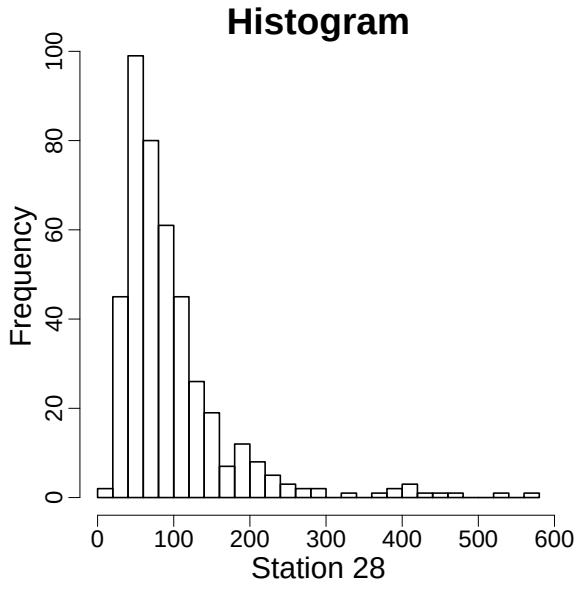


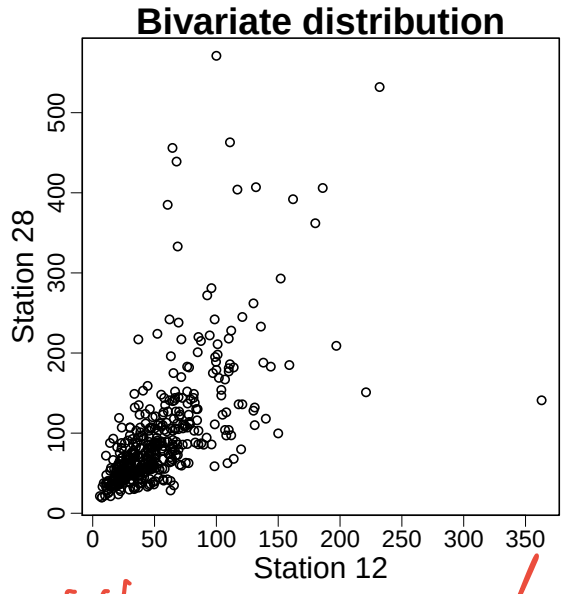
✗ it shows strong dependence

but hard to say!!!

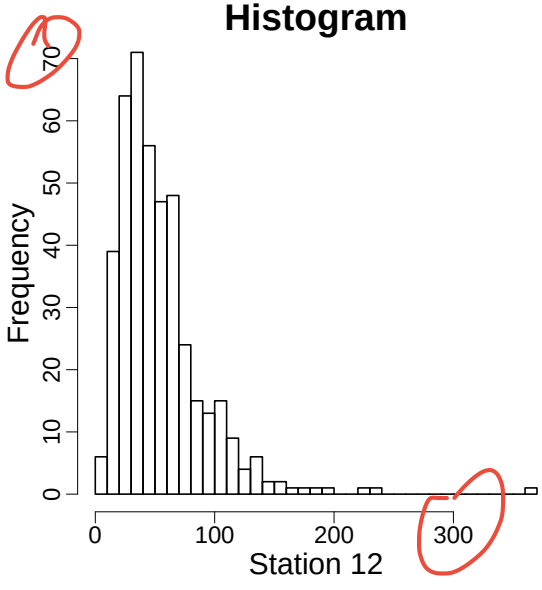
River discharge data

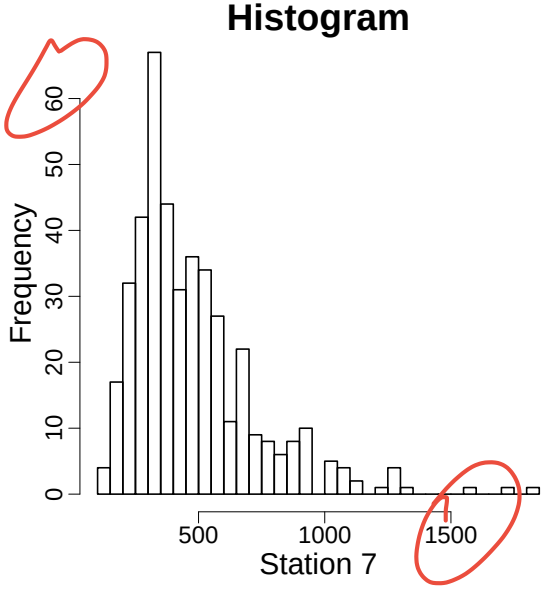


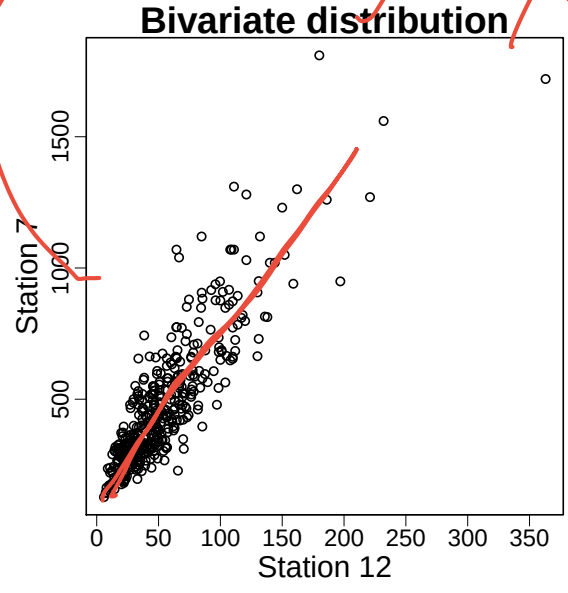




it's strong dependent







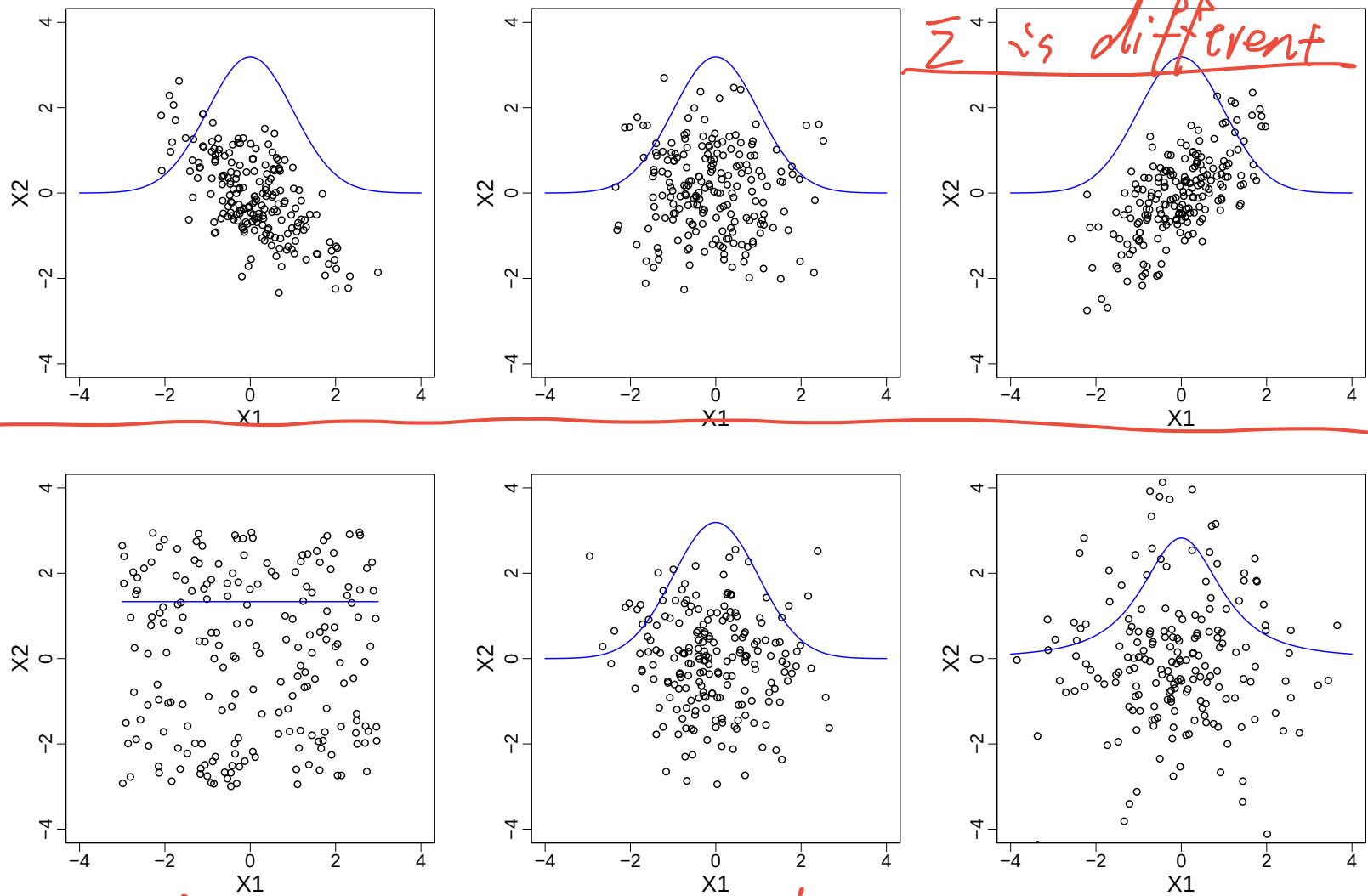
Simulated data

截面 xoy

they are same mean.

the marginal distribution is normal distribution

Σ is different



marginal → uniform

but they are independent

normal

T

Multivariate distributions

A multivariate random vector $(X_1, \ldots, X_p) \sim F$ has a distribution function

$$F(y_1, \ldots, y_p) = P(X_1 \leq y_1, \ldots, X_p \leq y_p),$$

and marginal distributions

$$F_j(y_j) = P(X_j \leq y_j) = F(\infty, \ldots, \infty, y_j, \infty, \ldots, \infty).$$

The idea of copulas is to **separate the marginals from the dependence**:

$$F \text{ “=” marginal cdfs } F_1, \ldots, F_p \text{ “+” dependence structure } C.$$

The normalized dependence structure is called the copula C.

Advantages

- We can understand and model the dependence independently of the margins.
- Model building and fitting is easier on the standardized level.
- There might be different amounts of information about margins and dependence.

Preliminaries: the quantile transformation

F can be not invertable

Let F be a univariate distribution function, the generalized inverse (also called quantile function) is

$$F^{\leftarrow}(p) = \inf\{x \in \mathbb{R} : F(x) \geq p\}.$$

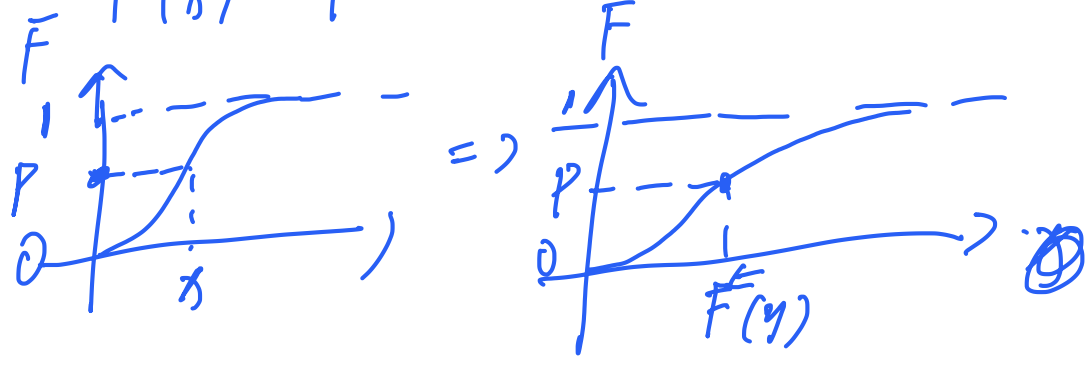
Note that if F is continuous then $F(F^{\leftarrow}(p)) = p$ and $F^{\leftarrow}(F(x)) = x$, for all $p \in [0,1]$ and $x \in \mathbb{R}$.

$F: \mathbb{R} \to [0,1]$

$F^{\leftarrow}: [0,1] \to \mathbb{R}$

exponential distribution

$F(x) = 1 - e^{-x}$ => CDF



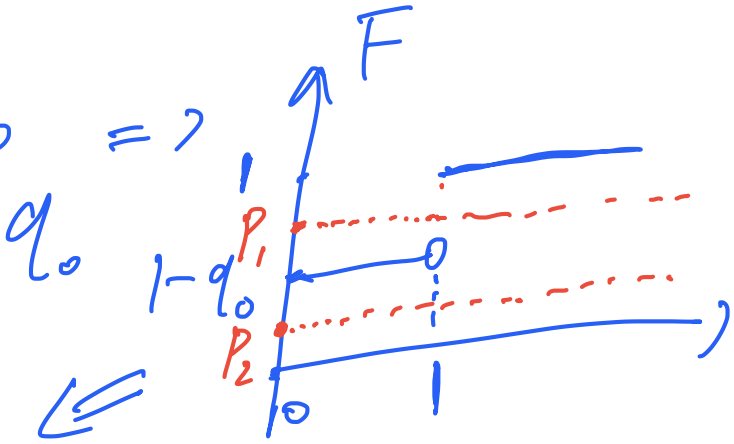
$F^{\leftarrow}(y) = -\log(1-p)$

bernouill distribution:

$Ber(q_0)$ $\quad q_0 \in [0,1]$

$P(X=1) = q_0 \Rightarrow$

$P(X=0) = 1-q_0$



$\Leftarrow$

$\overleftarrow{F}(p_1) = \inf\{x \in R : F(x) \geq p_1\}$

all ≥ 1 的 x 都满足.

$\overleftarrow{F}(p_2) = 0 \Leftarrow$ only 0 满足.

$$\overleftarrow{F}(p_2) = \begin{cases} 0, & \text{if } p \in [0, 1-q_0] \\ 1, & \text{if } p \in (1-q_0, 1] \end{cases}$$

Preliminaries: the quantile transformation

Let F be a univariate distribution function, the generalized inverse (also called quantile function) is

$$F^{\leftarrow}(p) = \inf\{x \in \mathbb{R} : F(x) \geq p\}.$$

Note that if F is continuous then $F(F^{\leftarrow}(p)) = p$ and $F^{\leftarrow}(F(x)) = x$, for all $p \in [0,1]$ and $x \in \mathbb{R}$.

Lemma

- Let $X \sim F$, F continuous, then $F(X) \sim U[0,1]$.
- If $U \sim U[0,1]$, then $F^{\leftarrow}(U) \sim F$.

Probability and quantile transformations are crucial to all applications involving copulas.

Examples:

- If $X \sim Ber(p_0)$, then $F(x) = (1-p_0)\mathbf{1}\{x \geq 0\} + p_0\mathbf{1}\{x \geq 1\}$, and $F^{\leftarrow}(U) = \mathbf{1}\{U > 1-p_0\} \sim Ber(p_0)$.
- If $X \sim \exp(\lambda)$ then $F(x) = 1 - \exp(-\lambda x)$ and $F(X) \sim U(0,1)$.

Definition of a copula

summary: $X \sim F \xrightarrow{F} U \sim U[0,1]$, $F^{\leftarrow}$ all sititution

Definition

A p-dimensional copula C is a multivariate distribution function $C : [0,1]^p \to [0,1]$ with standard uniform margins $U(0,1)$.

uniform margin of C

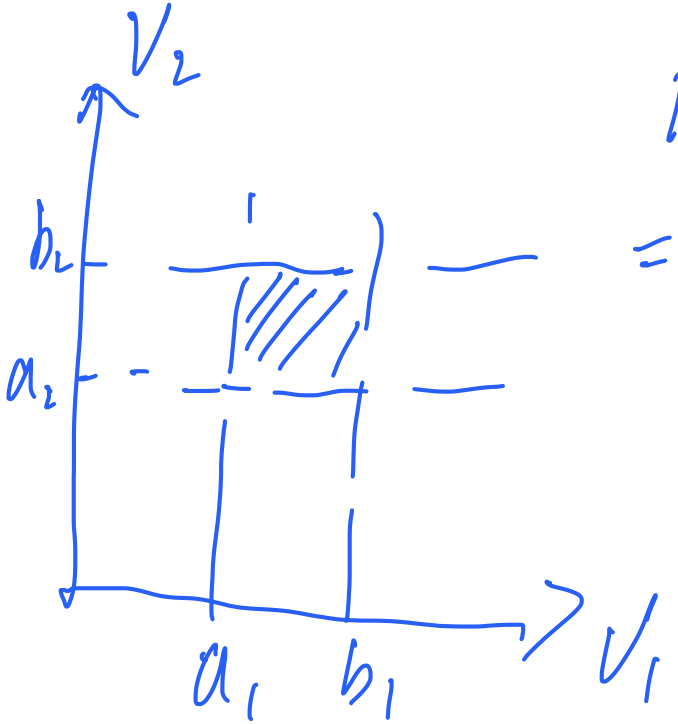
$C(1, 1, 1, x_i, \cdots 1) = x_i$

V_i is random

C is distribution

$V = (V_1 \cdots V_p) \sim C$

$P(V_1 \le 1, V_2 \le 1, V_i \le x_i, \cdots V_p \le 1) = x_i \in [0,1]$



$$P(V \in [a_1, b_1] \times [a_2, b_2]) = C(b_1, b_2) - C(b_1, a_2) - C(a_1, b_2) + C(a_1, a_2)$$

constrain: { 整个结果必在 $[0,1]$ A

其次每一项必在 $[0,1]$ B

copula 本质把一个多元积化为边缘乘把边缘用均匀替换. dependent 结构可用于均匀分布的联合表示, 即 copula.

A: P 必在 $[0,1]$

B: 首先构成 $C(b_1, b_2)$ 和边缘分布被转化为 均匀分布

我们要求这些分布在 $[0,1]$

Definition of a copula

Definition

A p-dimensional copula C is a multivariate distribution function $C : [0,1]^p \to [0,1]$ with standard uniform margins $U(0,1)$.

- A stochastic representation is $U \sim C$ with

$$P(U \leq u) = P(U_1 \leq u_1, \ldots, U_p \leq u_p) = C(u_1, \ldots, u_p),$$

- The margins of U are uniformly distributed

$$P(U_j \leq u_j) = C_j(u_j) = C(1, \ldots, u_j, \ldots, 1) = u_j.$$

- Any k-dimensional ($k < p$) marginal-subset is also a copula.

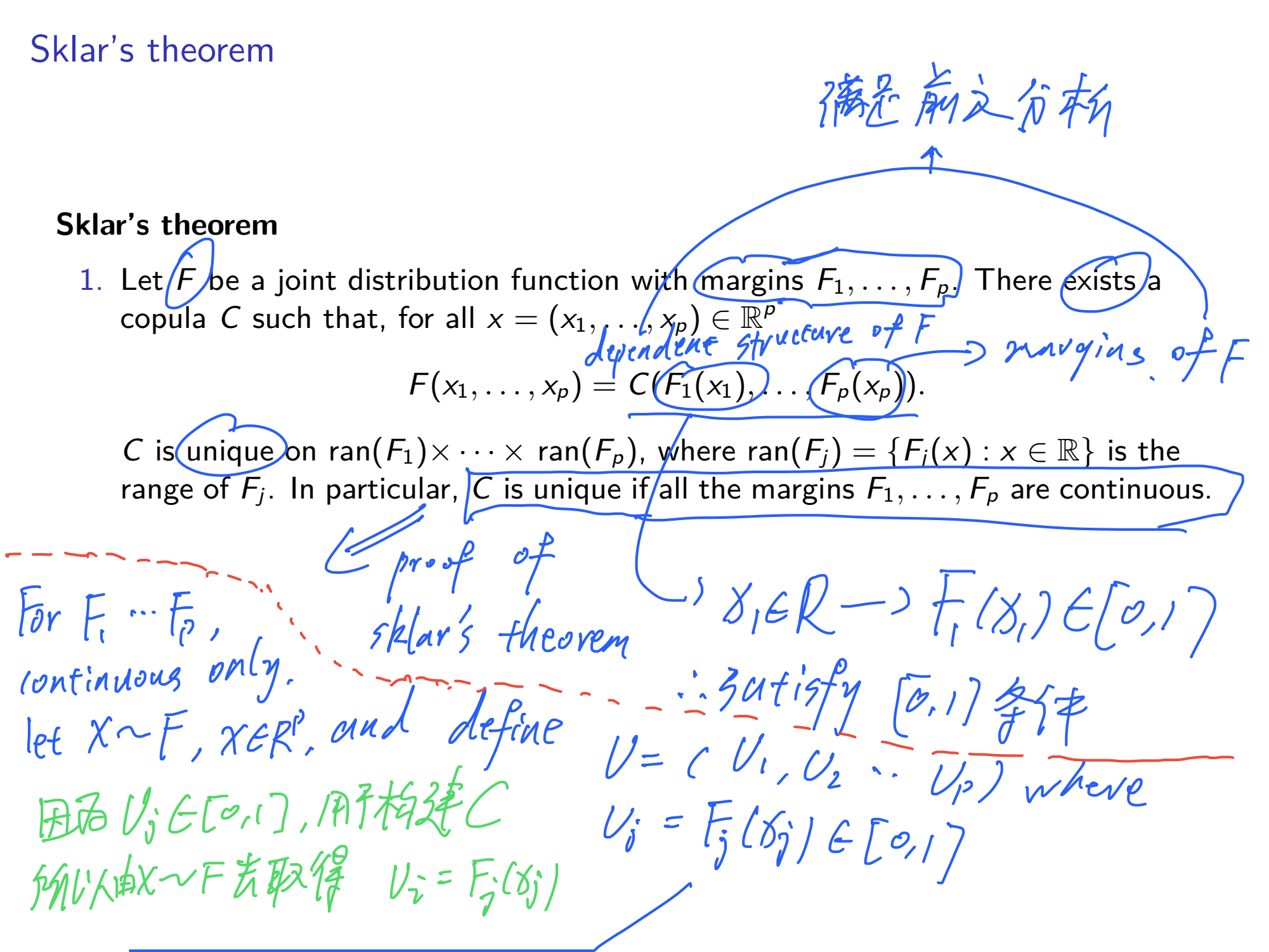
Sklar's theorem

Sklar's theorem

1. Let F be a joint distribution function with margins $F_1, \ldots, F_p$. There exists a copula C such that, for all $x = (x_1, \ldots, x_p) \in \mathbb{R}^p$

$$F(x_1, \ldots, x_p) = C(F_1(x_1), \ldots, F_p(x_p)).$$

C is unique on $\text{ran}(F_1) \times \cdots \times \text{ran}(F_p)$, where $\text{ran}(F_j) = \{F_j(x) : x \in \mathbb{R}\}$ is the range of F_j. In particular, C is unique if all the margins $F_1, \ldots, F_p$ are continuous.



$$F(x_1, x_2 \cdots x_p) = P(X_j \le x_j, \forall j) = P(F_j(X_j) \le F_j(x_j), \forall j)$$

$F_j(X_j) = U_j$

$$= P(U_j \le F_j(x_j), \forall j) = C(F_j(x_j), F_2(x_2), \ldots)$$

we define $U \sim C$ thus

Sklar's theorem

1. Let F be a joint distribution function with margins $F_1, \ldots, F_p$. There exists a copula C such that, for all $x = (x_1, \ldots, x_p) \in \mathbb{R}^p$

$$F(x_1, \ldots, x_p) = C(F_1(x_1), \ldots, F_p(x_p)).$$

   C is unique on $\text{ran}(F_1) \times \cdots \times \text{ran}(F_p)$, where $\text{ran}(F_j) = \{F_j(x) : x \in \mathbb{R}\}$ is the range of F_j. In particular, C is unique if all the margins $F_1, \ldots, F_p$ are continuous.

2. Conversely, if C is a copula and $F_1, \ldots, F_p$ are univariate distribution functions, then F defined as

$$F(x_1, \ldots, x_p) = C(F_1(x_1), \ldots, F_p(x_p))$$

   is a multivariate distribution function with margins $F_1, \ldots, F_p$.

and copula C

Sklar's theorem

If the margins of $X \sim F$, $F_1, \ldots, F_p$ are all continuous, then the (unique) copula of X is

$$C(u_1, \ldots, u_p) = F(F_1^{\leftarrow}(u_1), \ldots, F_p^{\leftarrow}(u_p)).$$

Thus, a distribution function F is equivalent to a copula C and its margins F_j. The copula fully characterizes the dependence between the components and we have

$$X \text{ has copula } C \Leftrightarrow (F_1(X_1), \ldots, F_p(X_p)) \sim C.$$

Sklar's theorem

If the margins of $X \sim F$, $F_1, \ldots, F_p$ are all continuous, then the (unique) copula of X is

$$C(u_1, \ldots, u_p) = F(F_1^{\leftarrow}(u_1), \ldots, F_p^{\leftarrow}(u_p)).$$

Thus, a distribution function F is equivalent to a copula C and its margins F_j. The copula fully characterizes the dependence between the components and we have

$$X \text{ has copula } C \Leftrightarrow (F_1(X_1), \ldots, F_p(X_p)) \sim C.$$

Invariance principle

Let $X \sim F$ with continuous margins $F_1, \ldots, F_p$ and copula C. Then for any strictly monotonically increasing T_j on $\text{ran}(X_j)$, we have that $(T_1(X_1), \ldots, T_p(X_p))$ has also copula C.

Examples

↳ 这也就说明了 边缘分布可同 ~~无法得~~.
但可能有 same copulas

if $T_j(x_j) = \lambda_j x_j$, $\lambda_j > 0$

$\lambda_j \leq 0$, C is unchanged

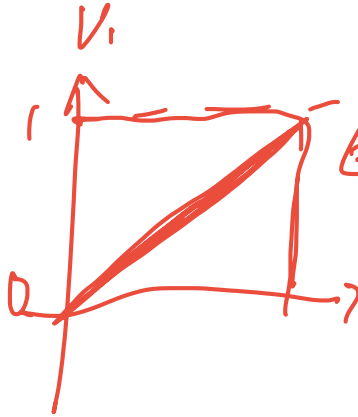
C is changed

Bivariate Bernoulli distribution

Let (X_1, X_2) follow a bivariate Bernoulli distribution with

$$P(X_1 = k, X_2 = l) = 1/4, k, l \in \{0, 1\},$$

then $X_1, X_2 \sim Ber(1/2)$. Any copula with $C(1/2, 1/2) = 1/4$ satisfies the equation in point 1. of Sklar's theorem on $\{0, 1/2, 1\} \times \{0, 1/2, 1\}$, e.g., $C(u_1, u_2) = u_1 u_2$.



now we changed it.

$\in \begin{cases} U_1 = U \sim \text{uniform}[0,1] \\ U_2 = U \sim \text{uniform}[0,1] \end{cases}$ $\vec{U} = (U_1, U_2)$

$X_1 = U$ $X_2 = U^2$ $\vec{X} = (X_1, X_2)$

$F_{X_1}(x_1) \sim \text{uniform}[0,1]$ $F_{X_2}(x_2) = P(X_2 \leq x)$ ⇐ marginal

$= P(U^2 \leq x)$ ∵ $U \sim \text{uniform}[0,1]$ 且 U 为随变

$= P(U \leq \sqrt{x}) = \sqrt{x} \in [0,1]$

$F_{X_2}(u) = u^2$

joint distribution (X_1, X_2) 与 (U_1, U_2) 因为 marginal different

$$C_{\vec{U}}(u_1, u_2) = P(U_1 \le u_1,\ U_2 \le u_2)$$

why ⇐ 因为 $U_1 = U$, $U_2 = U$

$$= P(U \le u_1,\ U \le u_2)$$

$$= P[U \le \min(u_1, u_2)] = \min(u_1, u_2)$$

$$C_{\vec{X}}(x_1, x_2) = P(X_1 \le F_{X_1}^{\leftarrow}(u_1),\ X_2 \le F_{X_2}^{\leftarrow}(u_2))$$ 用 ⑯ 的定义

$$= P(U \le u_1,\ U^2 \le u_2)$$ ← $x_1 = u$, $x_2 = u^2$

$$= P(U \le u_1,\ U \le u_2)$$

因为单调上升且>0 直接去.

$F_{X_1}^{\leftarrow}(u_1) = u_1$

$F_{X_2}^{\leftarrow}(u_2) = u_2^2$ 求逆

从 $F_X(x)$ 可得.

A 与 B 一样.

has the same Copulas.

summary $U\begin{cases} U_1 = U \\ U_2 = U \end{cases} \longrightarrow X\begin{cases} X_1 = U \Leftarrow T_1(X_1) = X_1 \\ X_2 = U^2 \Leftarrow T_2(X_2) = X_2^2 \end{cases}$.

---

let me another example:

$\vec{X}\begin{cases} X_1 = U. \\ X_2 = -U. \end{cases}$ $C_{\vec{X}}(\gamma_1, \gamma_2) = P(X_1 \le \overleftarrow{F}_{\gamma_1}(u_1),\ X_2 \le \overleftarrow{F}_{\gamma_2}(u_2))$

$= P(U \le u_1,\ -U \le u_2 - 1)$

$= P(U \le u_1,\ U \ge 1 - u_2)$

$= P(1 - u_2 \le U \le u_1) = P(u_1) - P(1 - u_2)$

$= u_1 - (1 - u_2)$

$= u_1 + u_2 - 1$

condition: $u_1 > 1 - u_2$

$\Downarrow$

$F_{\gamma_1}(\gamma_1) = u$.

$F_{\gamma_2}(\gamma_2) = -U$ X false.

true $= P(\gamma_2 \le X_2)$

$= P(-U \le \gamma_2)$

$= P(U \ge -\gamma_2)$

$= 1 - P(-\gamma_2) = 1 - (-\gamma_2) = 1 + \gamma_2$ $\begin{cases} 1 + \gamma_2 \in (0, 1) \\ \gamma_2 \in (-1, 0) \end{cases}$

$\overleftarrow{F}_{\gamma_2}(\gamma_2)$: let $\eta = 1 + \gamma_2$ $\longrightarrow \gamma_2 = \eta - 1$

$\overleftarrow{F}_{\gamma_2}(\gamma_2) = 1 - \gamma_2$

Examples

Bivariate Bernoulli distribution
Let (X_1, X_2) follow a bivariate Bernoulli distribution with

$$P(X_1 = k, X_2 = l) = 1/4, k, l \in \{0, 1\},$$

then $X_1, X_2 \sim Ber(1/2)$. Any copula with $C(1/2, 1/2) = 1/4$ satisfies the equation in point 1. of Sklar's theorem on $\{0, 1/2, 1\} \times \{0, 1/2, 1\}$, e.g., $C(u_1, u_2) = u_1 u_2$.

Gaussian copula
Let $\mathbf{R} \in \mathbb{R}^{p \times p}$ be a correlation matrix and Φ the univariate normal cdf, and Φ_R the p-dimensional cdf of a multivariate normal with mean zero and covariance matrix $\mathbf{R}$, then

$$C_{\mathbf{R}}(u_1, \ldots, u_p) = \Phi_{\mathbf{R}}(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_p))$$

is called the Gaussian copula with parameter matrix $\mathbf{R}$.

Examples

Independence copula

The $X_1, \ldots, X_p$ are independent if and only if their copula is $C(u) = \prod_{j=1}^{p} u_j$.

Examples

Independence copula
The $X_1, \ldots, X_p$ are independent if and only if their copula is $C(u) = \prod_{j=1}^{p} u_j$.
Countermonotonicity copula 负单调性 copula.
The countermonotonicity copula is the copula of $(U, 1-U)$ where $U \sim U[0,1]$. It can be shown that (X_1, X_2) has this copula if and only if they are perfectly negatively dependent, i.e., X_2 is a strictly decreasing function of X_1.

Examples

Independence copula

The $X_1, \ldots, X_p$ are independent if and only if their copula is $C(u) = \prod_{j=1}^{p} u_j$.

Countermonotonicity copula

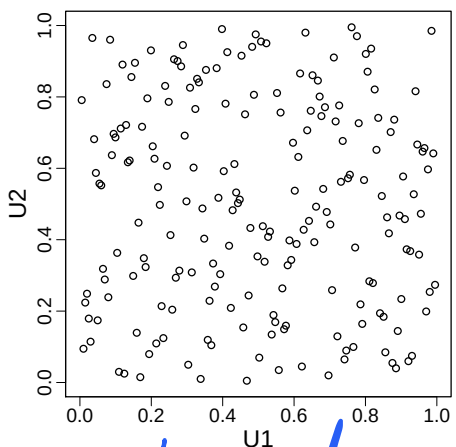
The countermonotonicity copula is the copula of $(U, 1-U)$ where $U \sim U[0,1]$. It can be shown that (X_1, X_2) has this copula if and only if they are perfectly negatively dependent, i.e., X_2 is a strictly decreasing function of X_1.

Comonotonicity copula

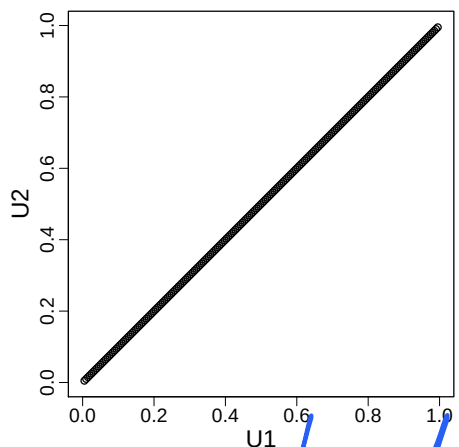
The comonotonicity copula is the copula of $(U, \ldots, U)$ where $U \sim U[0,1]$. It can be shown that $(X_1, \ldots, X_p)$ has this copula if and only if the components are perfectly positively dependent, i.e., $X_2, \ldots, X_p$ are a strictly increasing functions of X_1.

Examples

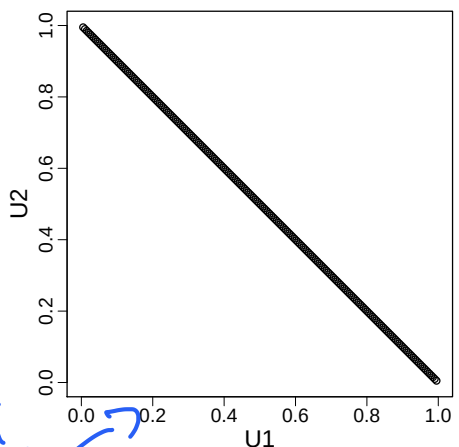
Independence copula
The $X_1, \ldots, X_p$ are independent if and only if their copula is $C(u) = \prod_{j=1}^{p} u_j$.
Countermonotonicity copula
The countermonotonicity copula is the copula of $(U, 1-U)$ where $U \sim U[0,1]$. It can be shown that (X_1, X_2) has this copula if and only if they are perfectly negatively dependent, i.e., X_2 is a strictly decreasing function of X_1.
Comonotonicity copula 共同性 copula.
The comonotonicity copula is the copula of $(U, \ldots, U)$ where $U \sim U[0,1]$. It can be shown that $(X_1, \ldots, X_p)$ has this copula if and only if the components are perfectly positively dependent, i.e., $X_2, \ldots, X_p$ are a strictly increasing functions of X_1.



independent

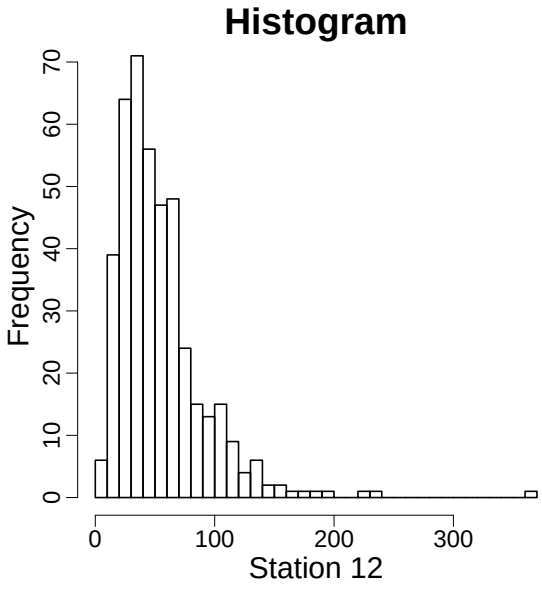


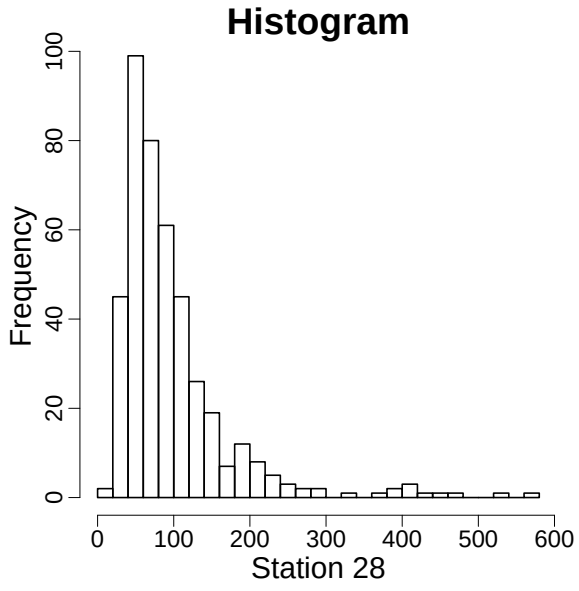
strong dependent
comonotoicity

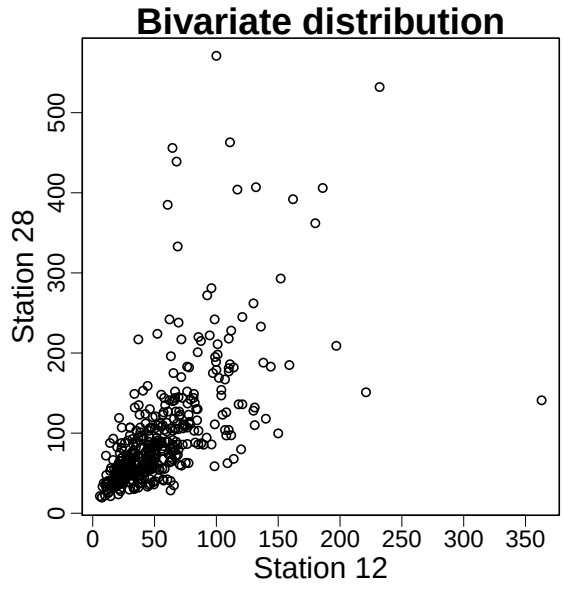


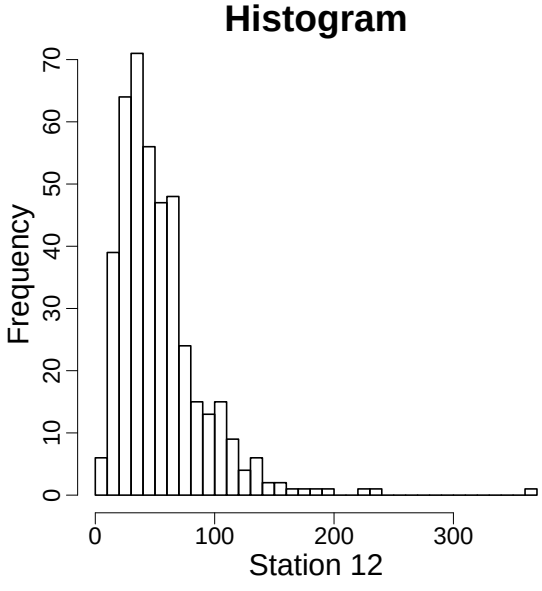
countermonotonicity

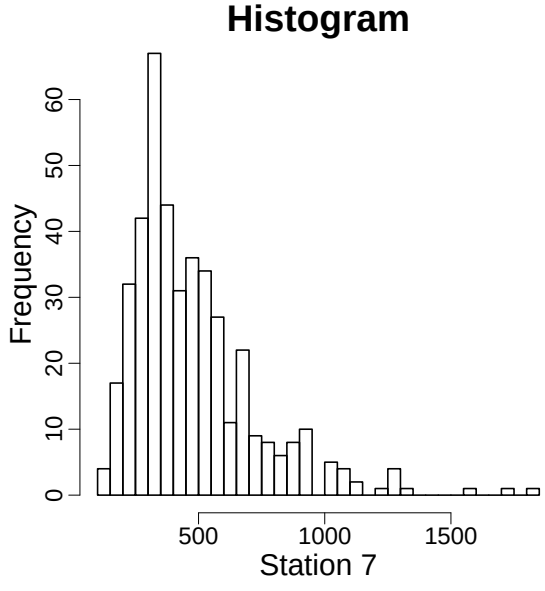
River discharge data

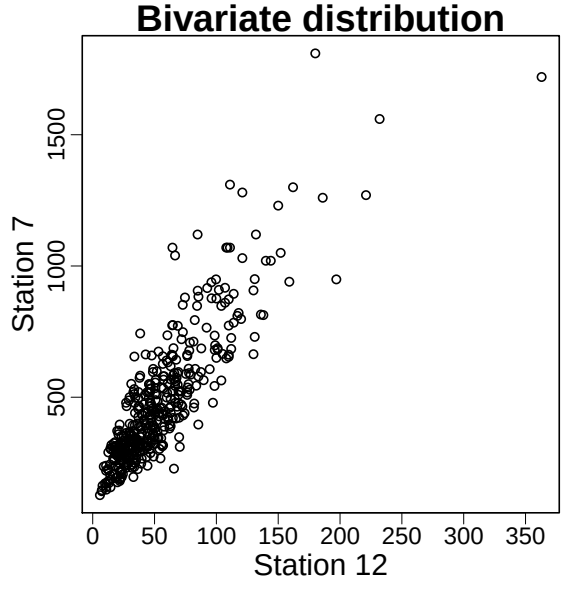




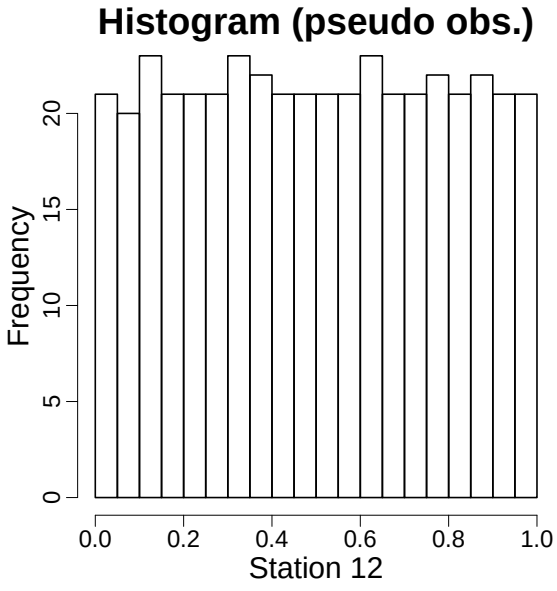


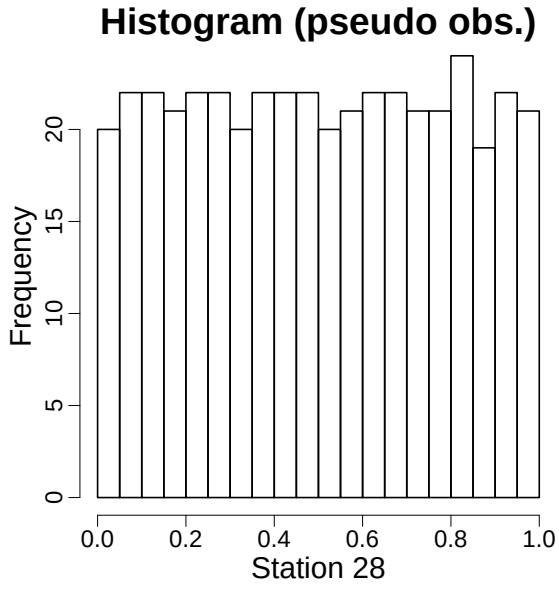


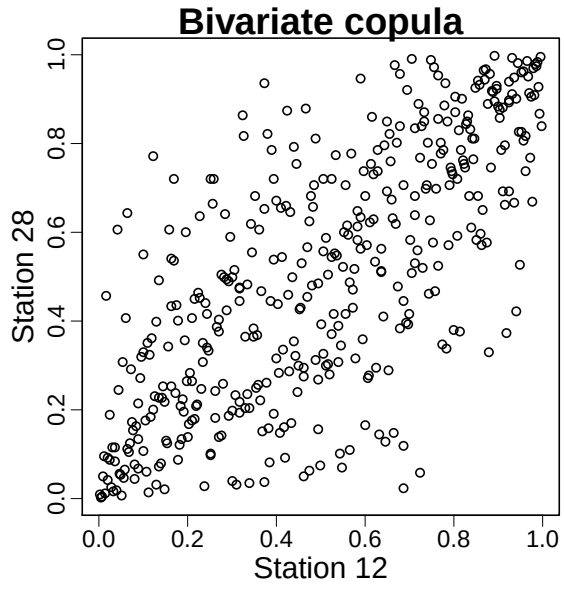


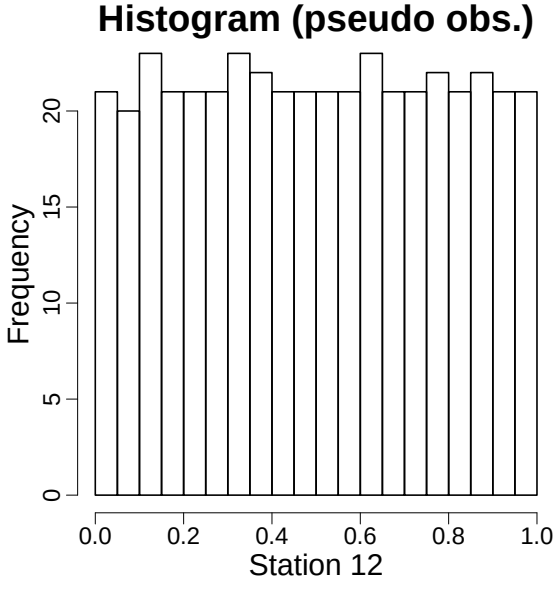


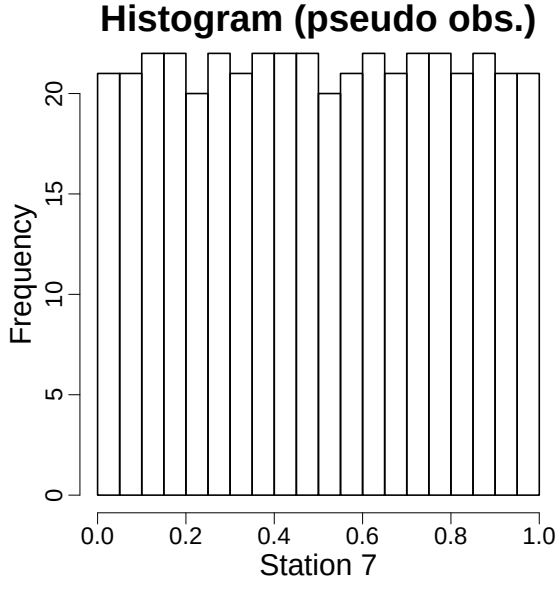
River discharge data

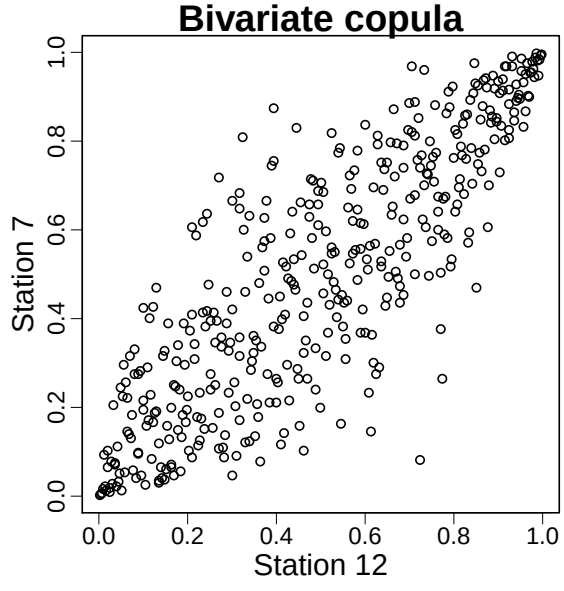




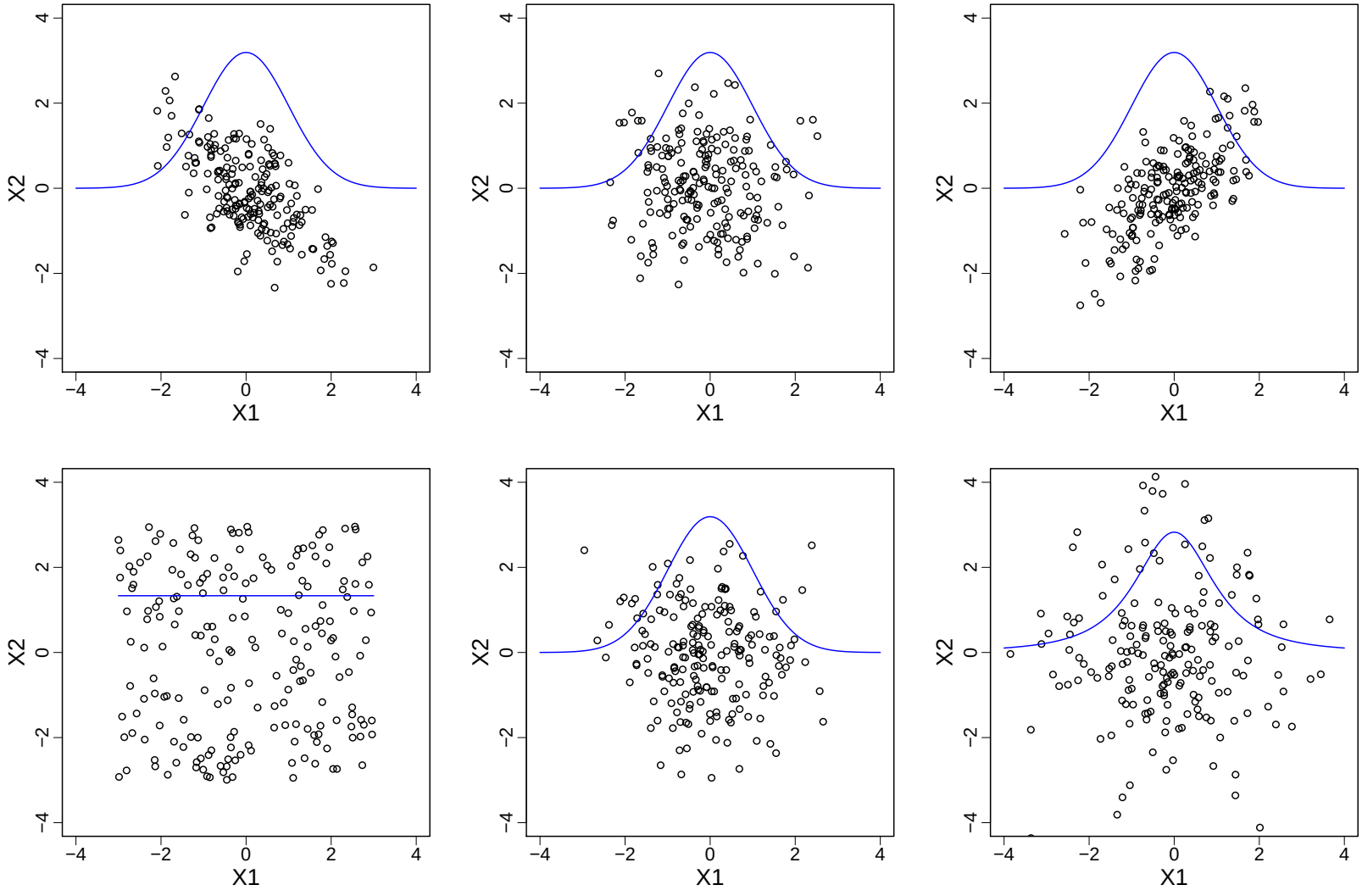




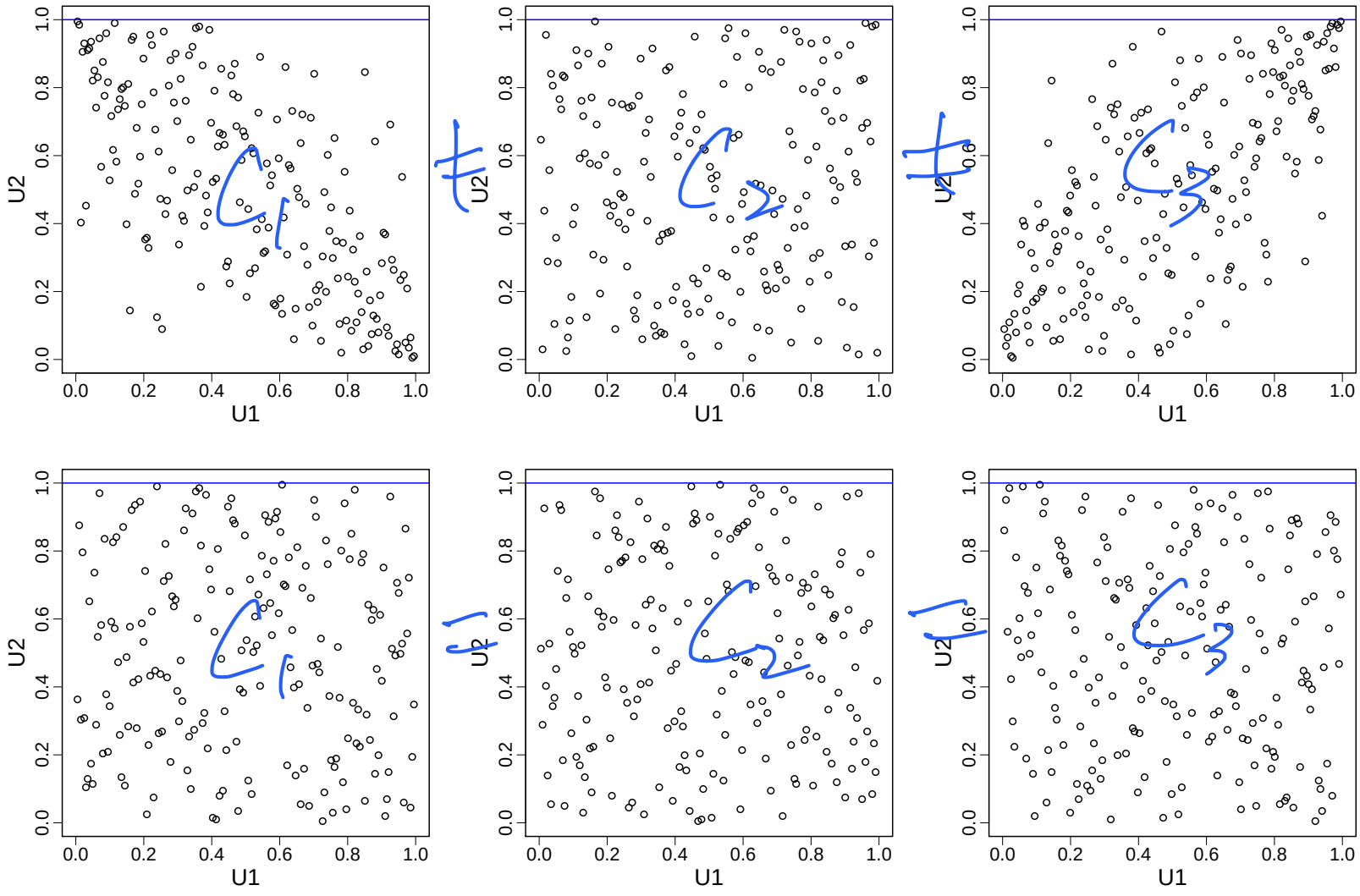




Simulated data



Simulated data



Fréchet–Höffding bounds

Theorem (Fréchet–Höffding bounds)
Let $W(u) = \max\{\sum_{j=1}^{p} u_j - p + 1, 0\}$ and $M(u) = \min_{j=1,\dots,p} u_j$.

(1) For any p-dimensional copula C

$$W(u) \le C(u) \le M(u), \quad u \in [0,1]^p.$$

(2) W is a copula if and only if $p = 2$.

(3) M is a copula for all $p \ge 2$

Note:
It is easy to see that for $U \sim U[0,1]$

- $(U, \dots, U) \sim M$ for any p.
- $(U, 1-U) \sim W$ for $p = 2$.

Proof of (2):

$W(1, \cdots u_j \cdots 1) = \max\{p-1+u_j+1-p, 0\}$
$= \max\{u_j, 0\}$

$u_j \in [0,1]$
all marginal distribution uniform.

M we also don't show but we can construct it.

but the key problem is that W is only function we not show it as joint CDF

Densities of copulas

- Let $(X_1, \ldots, X_p)$ have differentiable cdf F with joint and marginal densities given by f and $f_1, \ldots, f_p$, respectively. Then the corresponding copula has density

$$c(u_1, \ldots, u_p) = \frac{\partial^p C(u_1, \ldots, u_p)}{\partial u_1 \ldots \partial u_p}$$

$$= \frac{f\left(F_1^{-1}(u_1), \ldots, F_p^{-1}(u_p)\right)}{\prod_{j=1}^{p} f_j\left(F_j^{-1}(u_j)\right)}$$

Example:
The Gaussian copula has density

$$c_{\mathbf{R}}(u_1, \ldots, u_p) = \frac{1}{\sqrt{\det \mathbf{R}}} \exp\left(-\frac{1}{2} x^T (\mathbf{R}^{-1} - I_p) x\right),$$

where $x = (\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_p))$.

$X \sim N(0, R)$

$$f_X(x) = \frac{1}{\sqrt{2\pi} \cdot \sqrt{\det(R)}^T} \cdot \exp\left(-\frac{1}{2} x^T R^{-1} x\right)$$

$$f_1(x_1) \cdots f_p(x_p) = \frac{1}{(2\pi)^{p/2}} \cdot \exp \cdot \left(-\frac{1}{2} x^T x\right)$$

$$\text{分母} = \frac{1}{(2\pi)^{p/2}} \exp \cdot \left(-\frac{1}{2} x^T \cdot I_p x\right)$$

Archimedean copulas

Archimedean copulas are copulas of the form

$$C(u) = \psi(\psi^{-1}(u_1) + \cdots + \psi^{-1}(u_p)),$$

where ψ is called the (Archimedean) generator, where

- $\psi : [0, \infty) \to [0, 1]$ is strictly decreasing on its support and it satisfies $\psi(0) = 1$ and $\psi(\infty) = \lim_{t\to\infty} \psi(t) = 0$.
- There are further conditions to be satisfied in order that ψ generates a proper copula (e.g., complete monotonicity).

Archimedean copulas: Gumbel and Clayton

Taking particular (admissible) functions ψ we can construct particular examples of Archimedean copulas.

Gumbel copula

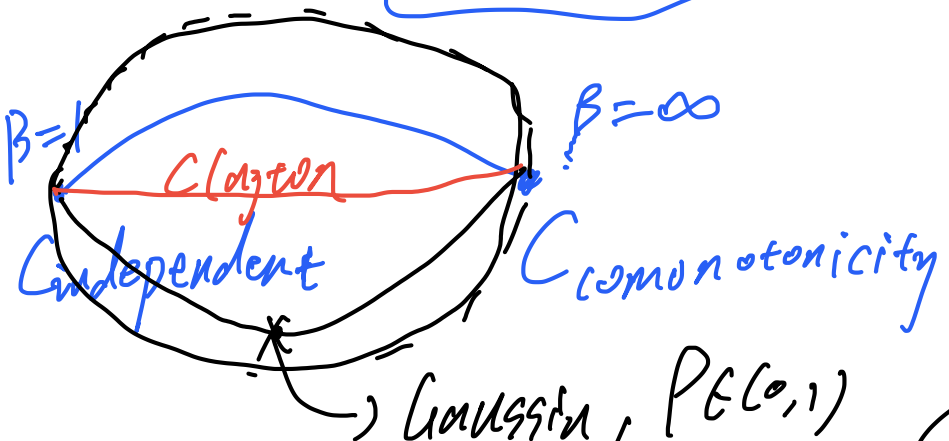
- Let $\psi(t) = \exp(-t^{1/\beta})$, for $\beta \geq 1$, then use the define

$$C_\beta^{Gu}(u_1, \ldots, u_p) = \exp\left(-\left((-\log u_1)^\beta + \cdots + (-\log u_p)^\beta\right)^{1/\beta}\right).$$

- $\beta = 1$ gives independence since

$$C_\beta^{Gu}(u_1, \ldots, u_p) = \exp\left(\log u_1 \cdots + \log u_p\right) = \prod_{j=1}^{p} u_j$$

- $\beta \to \infty$ gives comonotonicity.



but we cannot get all the copulas. from independence to dependence!!! many all copulas is like O, we only get blue by Gumbel

Archimedean copulas: Gumbel and Clayton

Taking particular (admissible) functions ψ we can construct particular examples of Archimedean copulas.

Gumbel copula

- Let $\psi(t) = \exp(-t^{1/\beta})$, for $\beta \geq 1$, then

$$C_\beta^{Gu}(u_1, \ldots, u_p) = \exp\left(-\left((-\log u_1)^\beta + \cdots + (-\log u_p)^\beta\right)^{1/\beta}\right).$$

- $\beta = 1$ gives independence since

$$C_\beta^{Gu}(u_1, \ldots, u_p) = \exp\left(\log u_1 \cdots + \log u_p\right) = \prod_{j=1}^{p} u_j$$

- $\beta \to \infty$ gives comonotonicity.

Clayton copula

- Let $\psi(t) = (1+t)^{-1/\beta}$, for $\beta \in (0, \infty)$, then

$$C_\beta^{Cl}(u_1, \ldots, u_p) = \left(u_1^{-\beta} + \cdots + u_p^{-\beta} - p + 1\right)^{-1/\beta}.$$

- $\beta = 0$ gives independence and $\beta \to \infty$ gives comonotonicity.

Archimedean copulas: properties

Advantages

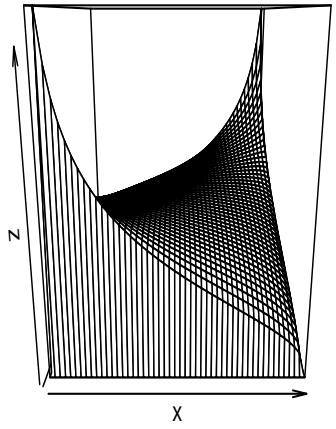
- Typically they are explicit (if ψ^{-1} is available).
- Properties can be expressed in terms of the generator ψ.
- Densities are available in many examples.
- Simulation is often simple.
- Flexible: not restricted to radial symmetry.

Drawbacks

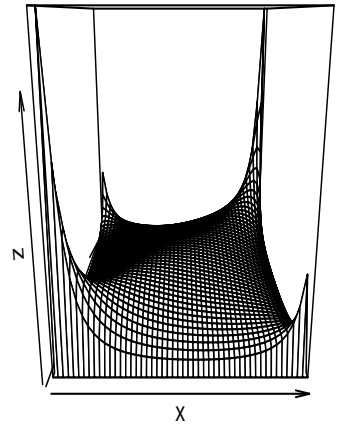
- All d-dimensional margins, $d < p$, have the same distribution.
- Often used only with few parameters, much less than $p(p-1)/2$ as in the Gaussian case.

density of different copulas

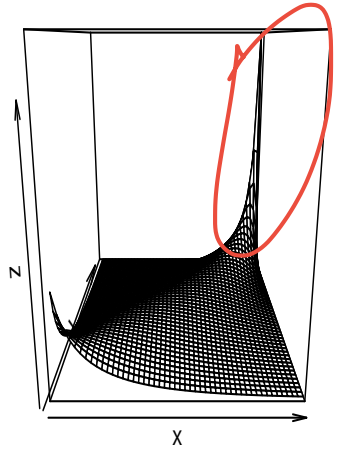
Gaussian, r = 0.3



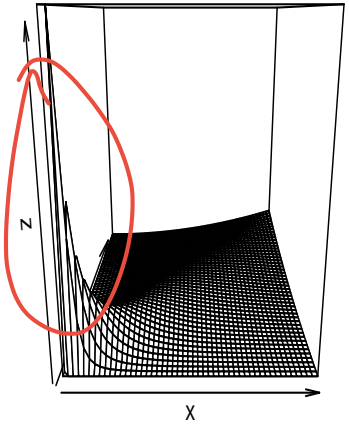
Student, r = 0.3, df = 2



Gumbel, beta = 2

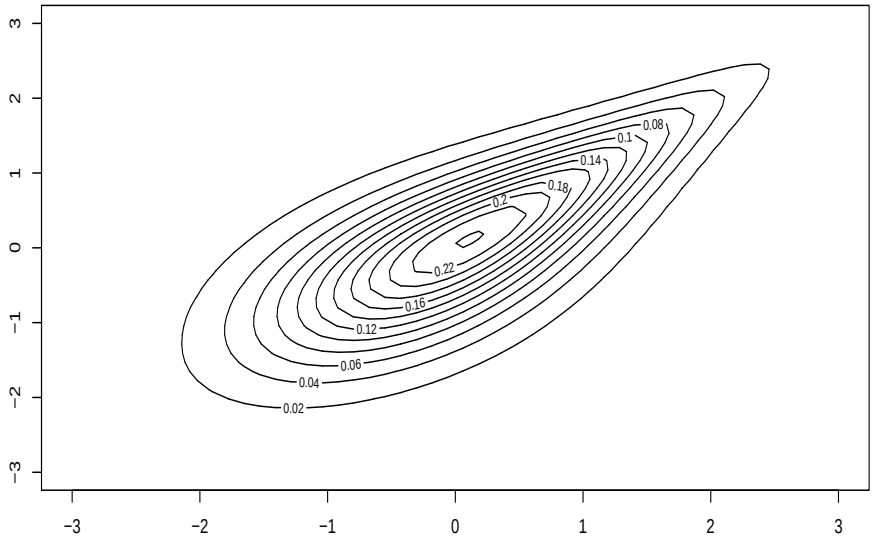


Clayton, beta = 2

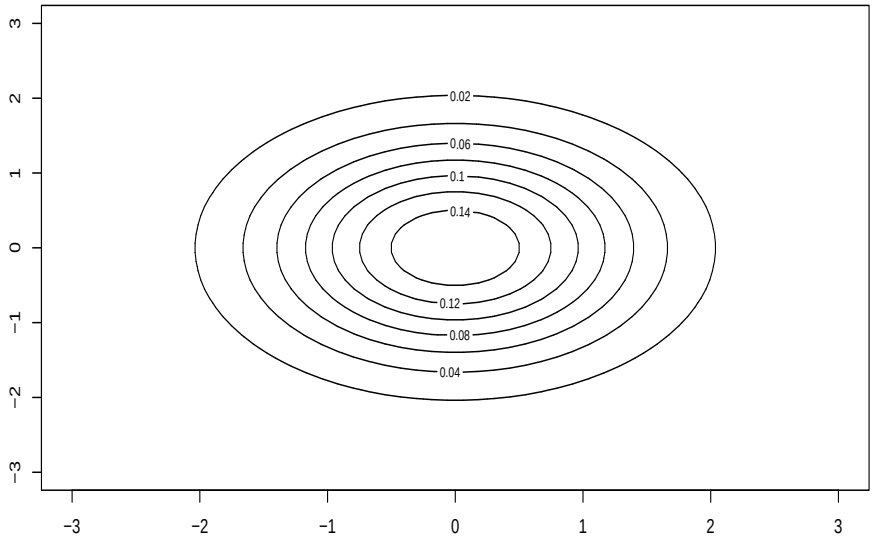


XY 投影

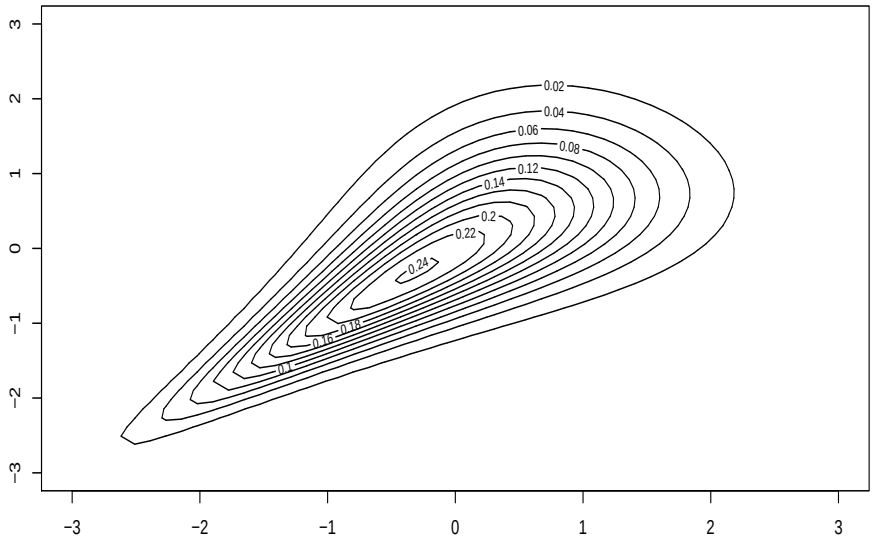
Gumbel on Normal marginal, beta = 2



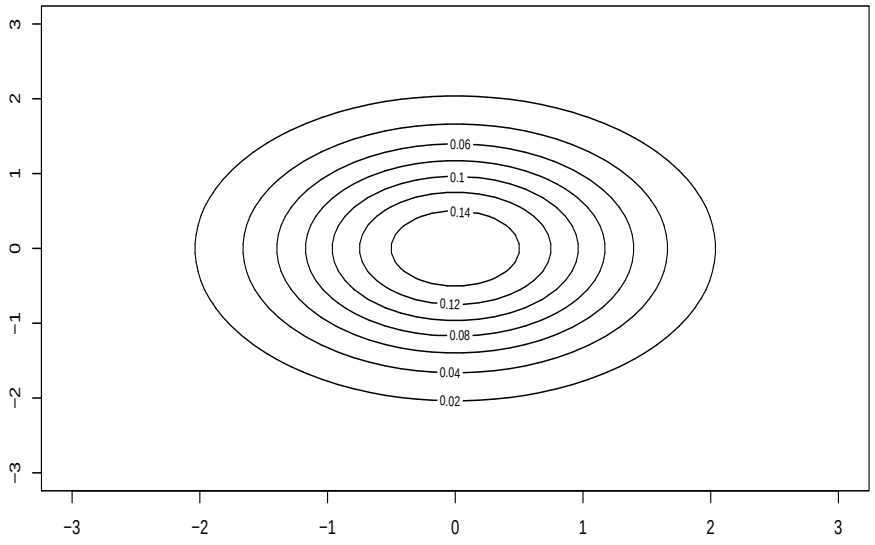
Gumbel on Normal marginal, beta = 1



Clayton on Normal marginal, beta = 2



Clayton on Normal marginal, beta = 0.00001



Statistics with copulas

- Suppose we have data $X_1, \ldots, X_n \in \mathbb{R}^p$ from a distribution F with copula C.
- In order to apply the copula framework, we first need to model the univariate marginal distributions with estimators $\widehat{F}_j$ of F_j, $j = 1, \ldots, p$.
- We can then introduce **pseudo-observations**

$$\widehat{U}_i = (\widehat{U}_{i1}, \ldots, \widehat{U}_{ip}) = (\widehat{F}_1(X_{i1}), \ldots, \widehat{F}_p(X_{ip})) \qquad i = 1, \ldots, n.$$

- Note that the $\widehat{U}_{ij}$ are only approximately $U[0,1]$ distributed.

因为是估计 所以是 approximately.

所以问题转变为如何估计 $\hat{F}_j$ from data.

Statistics with copulas

- Suppose we have data $X_1, \ldots, X_n \in \mathbb{R}^p$ from a distribution F with copula C.
- In order to apply the copula framework, we first need to model the univariate marginal distributions with estimators $\widehat{F}_j$ of F_j, $j = 1, \ldots, p$.
- We can then introduce **pseudo-observations**

$$\widehat{U}_i = (\widehat{U}_{i1}, \ldots, \widehat{U}_{ip}) = (\widehat{F}_1(X_{i1}), \ldots, \widehat{F}_p(X_{ip})) \qquad i = 1, \ldots, n.$$

- Note that the $\widehat{U}_{ij}$ are only approximately $U[0,1]$ distributed.

Possible choices for $\widehat{F}_j$:

- Parametric estimator using a parametric family, e.g. Gaussian or t-distribution.

Statistics with copulas

- Suppose we have data $X_1, \dots, X_n \in \mathbb{R}^p$ from a distribution F with copula C.
- In order to apply the copula framework, we first need to model the univariate marginal distributions with estimators $\widehat{F}_j$ of F_j, $j = 1, \dots, p$.
- We can then introduce **pseudo-observations**

$$\widehat{U}_i = (\widehat{U}_{i1}, \dots, \widehat{U}_{ip}) = (\widehat{F}_1(X_{i1}), \dots, \widehat{F}_p(X_{ip})) \qquad i = 1, \dots, n.$$

- Note that the $\widehat{U}_{ij}$ are only approximately $U[0,1]$ distributed.

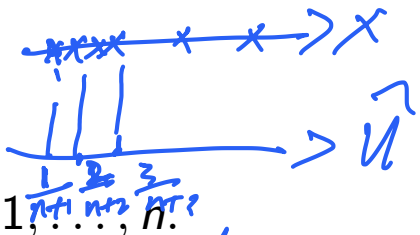
Possible choices for $\widehat{F}_j$:

- Parametric estimator using a parametric family, e.g. Gaussian or t-distribution.
- Semi-parametric estimators, e.g. based on extreme value theory where the bulk of the distribution is modeled empirically and the tail is approximated by a so-called generalized Pareto distribution.

Statistics with copulas

- Suppose we have data $X_1, \dots, X_n \in \mathbb{R}^p$ from a distribution F with copula C.
- In order to apply the copula framework, we first need to model the univariate marginal distributions with estimators $\widehat{F}_j$ of F_j, $j = 1, \dots, p$.
- We can then introduce **pseudo-observations**

$$\widehat{U}_i = (\widehat{U}_{i1}, \dots, \widehat{U}_{ip}) = (\widehat{F}_1(X_{i1}), \dots, \widehat{F}_p(X_{ip})) \qquad i = 1, \dots, n.$$



- Note that the $\widehat{U}_{ij}$ are only approximately $U[0,1]$ distributed.

$$\widehat{F}_j(x) = \frac{1}{n} \sum_{i=1}^{n} \{X_{ij} \le x\}$$

Possible choices for $\widehat{F}_j$:

- Parametric estimator using a parametric family, e.g. Gaussian or t-distribution.
- Semi-parametric estimators, e.g. based on extreme value theory where the bulk of the distribution is modeled empirically and the tail is approximated by a so-called generalized Pareto distribution.
- Non-parametric estimators using the empirical cdf

$$\widehat{U}_{ij} = \frac{n}{n+1} \widehat{F}_{n,j}(X_{ij}) = \frac{R_{ij}}{n+1},$$

where R_{ij} is the rank of X_{ij} among all $X_{1j}, \dots, X_{nj}$.

Indicator function.

MC or 蒙特卡洛仿真

要排序的！排序后的0,1,2,3提前

Statistics with copulas

Estimation

- The log-likelihood of the data $X_1, \ldots, X_n \in \mathbb{R}^p$ from a distribution F_θ with copula C_{θ_C} is

$$\ell(\theta; X_1, \ldots, X_n) = \sum_{i=1}^{n} \ell(\theta, X_i)$$

$$= \sum_{i=1}^{n} \ell_C(\theta_C; F_1(X_{i1}, \theta_1), \ldots, F_p(X_{ip}, \theta_p)) + \sum_{i=1}^{n} \sum_{j=1}^{p} \ell_j(\theta_j; X_{ij}),$$

where

$$\ell_C(\theta_C; u_1, \ldots, u_p) = \log c(u_1, \ldots, u_p; \theta_C)$$

$$\ell_j(\theta_j; x) = \log f_j(x; \theta_j), \quad j = 1, \ldots, p.$$

and the θ_C are the copula parameters and the θ_j the marginal parameters.

Statistics with copulas

Estimation

- The log-likelihood of the data $X_1, \ldots, X_n \in \mathbb{R}^p$ from a distribution F_θ with copula C_{θ_C} is

$$\begin{aligned}\ell(\theta; X_1, \ldots, X_n) &= \sum_{i=1}^{n} \ell(\theta, X_i) \\ &= \sum_{i=1}^{n} \ell_C(\theta_C; F_1(X_{i1}, \theta_1), \ldots, F_p(X_{ip}, \theta_p)) + \sum_{i=1}^{n}\sum_{j=1}^{p} \ell_j(\theta_j; X_{ij}),\end{aligned}$$

  where

$$\begin{aligned}\ell_C(\theta_C; u_1, \ldots, u_p) &= \log c(u_1, \ldots, u_p; \theta_C) \\ \ell_j(\theta_j; x) &= \log f_j(x; \theta_j), \quad j = 1, \ldots, p.\end{aligned}$$

  and the θ_C are the copula parameters and the θ_j the marginal parameters.

This suggests a **two-step estimation** procedure:

- First estimate the marginal parameters θ_j separately (or non-parameterically);
- and then use the pseudo-observations to estimate θ_C by maximizing

$$\sum_{i=1}^{n} \ell_C(\theta_C; \widehat{U}_{i1}, \ldots, \widehat{U}_{ip})$$

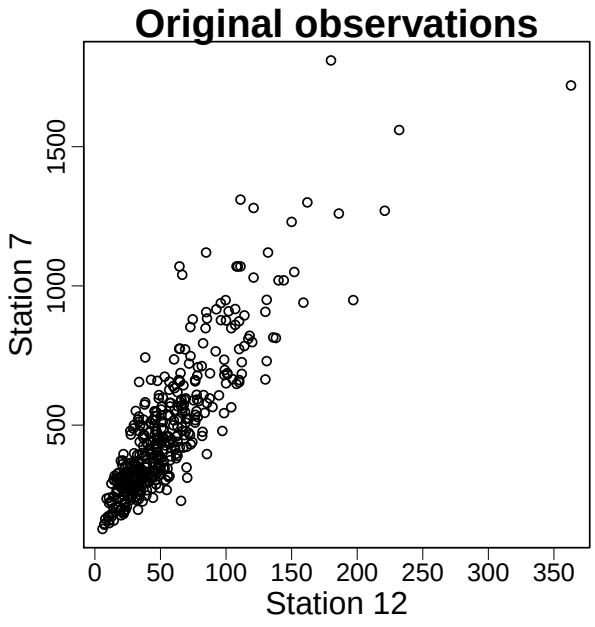
River discharge data: fitting in R

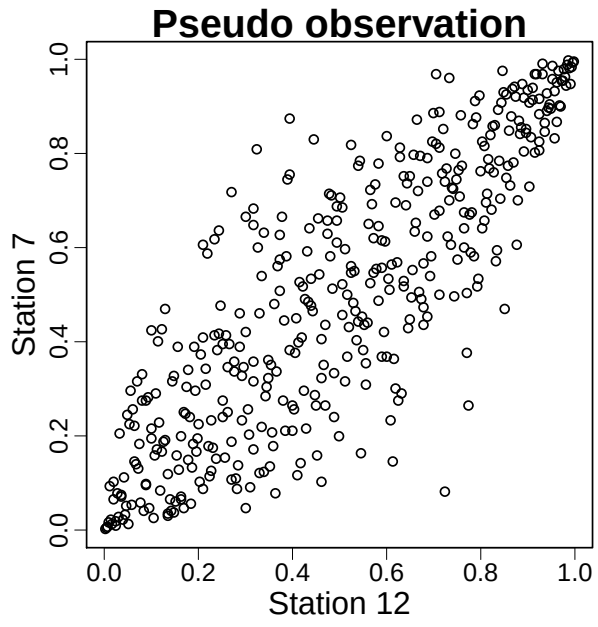
We transform the data non-parameterically and then fit a Gaussian copula. We can simulate from the copula and on the original scale.

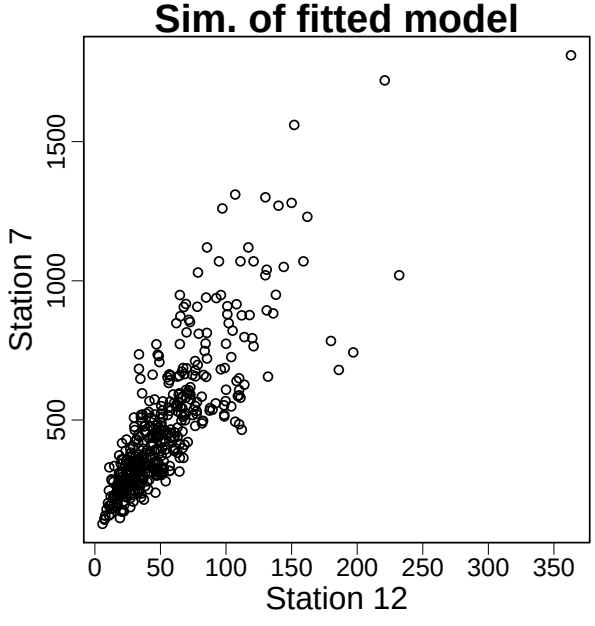
```
> library(copula)
> dim(X)
[1] 428  31
> U <- pobs(X)
> i <- 12
> j <- 7
> plot(X[,c(i,j)])
> plot(U[,c(i,j)])
> cop.fit <- fitCopula(normalCopula(), U[,c(i,j)], method="ml")
> cop.fit
fitCopula() estimation based on 'maximum likelihood'
and a sample of size 428.
      Estimate Std. Error z value Pr(>|z|)
rho.1 0.863887   0.009297   92.92   <2e-16 ***
> Usim <- rCopula(n = nrow(U), copula = cop.fit@copula)
> Xsim <- cbind(sort(X[,i])[rank(Usim[,1])], sort(X[,j])[rank(Usim[,2])])
> plot(Xsim)
> plot(Usim)
```

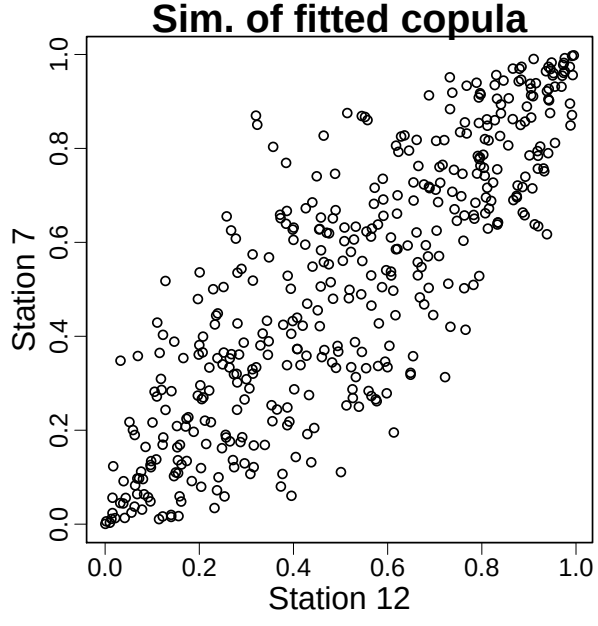
$n = 428 \quad p = 31$

River discharge data

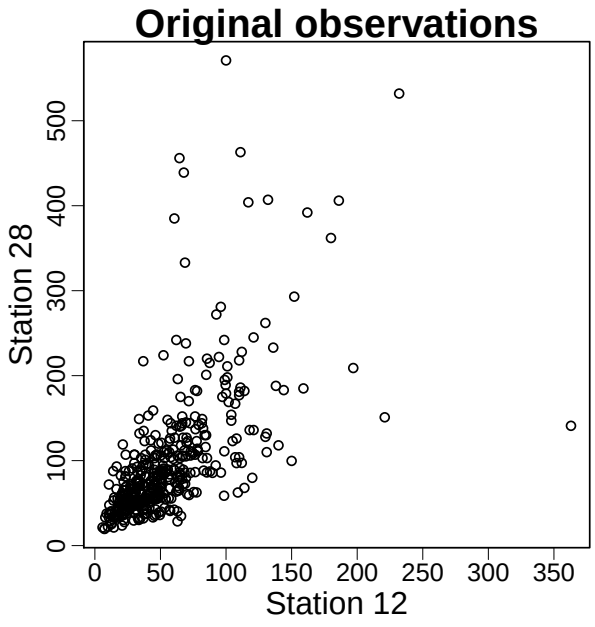


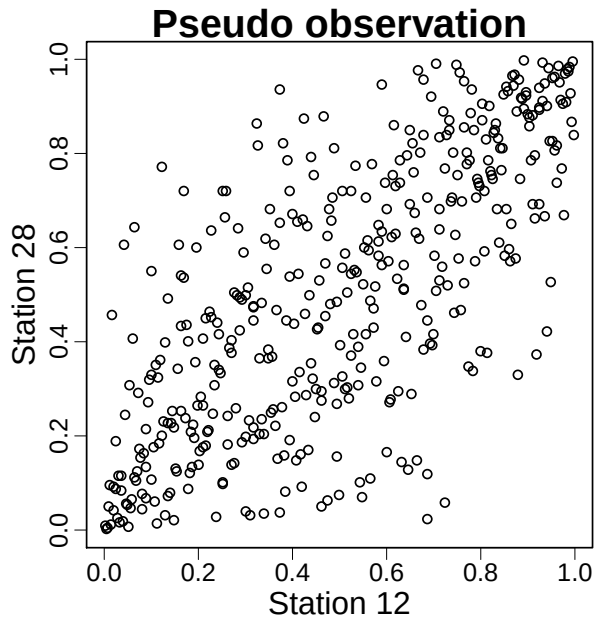


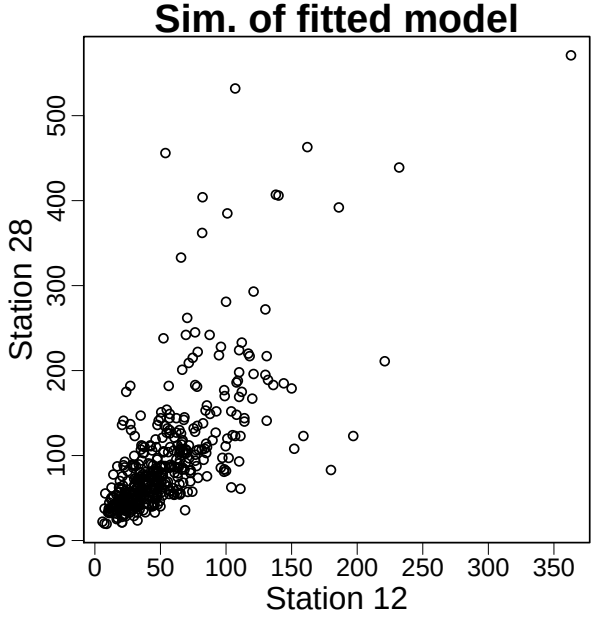


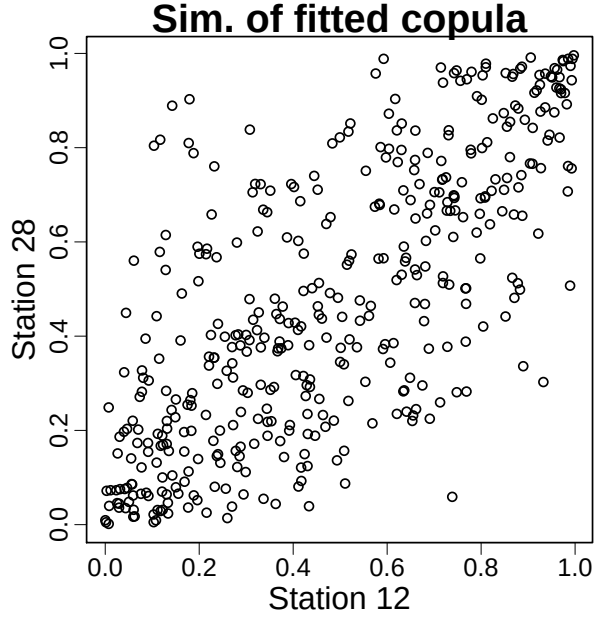


River discharge data









Measures of dependence: linear correlation

- Measures of dependence summarize a complicated dependence structure in a single number (in the bivariate case) or a matrix.
- For two random variables X_1 and X_2 with $E(X_j^2) < \infty$ the (Pearson) correlation coefficient is defined as

$$\rho(X_1, X_2) = \frac{\mathrm{cov}(X_1, X_2)}{\sqrt{\mathrm{Var} X_1}\sqrt{\mathrm{Var} X_2}}.$$

disadv: only linear

adv: put data in original scale

ρ is invariance for distrib

Properties of Pearson correlation

Some properties:

- $|\rho(X_1, X_2)| \leq 1$. Moreover, $|\rho(X_1, X_2)| = 1$ if and only if $X_2 = aX_1 + b$ almost surely for constants $a \neq 0$, $b \in \mathbb{R}$. This discards other strong functional dependence such as $X_2 = X_1^2$ for instance.
- X_1, X_2 independent $\Rightarrow \rho(X_1, X_2) = 0$. But $\rho(X_1, X_2) = 0 \not\Rightarrow X_1, X_2$ independent except for the normal distribution.
- "Marginal distributions and correlation determine the joint distribution": This is true for the class of multivariate normal distributions.
- $\rho(X_1, X_2)$ is invariant under strictly increasing `linear` transformations, but not under strictly increasing functions in general.
- For a class of elliptical distributions (e.g., normal or t), the correlation matrix together with the marginal distributions determine the whole distribution.

Alternative measures based on the ranks: rank correlations.

Spearman's rho

- Let $X = (X_1, X_2)$ be a random vector with distribution function F and copula C.
- The Spearman's rho between X_1 and X_2 is

$$\rho_S(X_1, X_2) = \rho(F_1(X_1), F_2(X_2)),$$

where ρ is the Pearson's correlation.

Spearman's rho

- Let $X = (X_1, X_2)$ be a random vector with distribution function F and copula C.
- The Spearman's rho between X_1 and X_2 is

$$\rho_S(X_1, X_2) = \rho(F_1(X_1), F_2(X_2)),$$

  where ρ is the Pearson's correlation.
- One can show that this is equivalent to → exercise I you can see it

$$\rho_S(X_1, X_2) = 12 \int_0^1 \int_0^1 \{C(u_1, u_2) - u_1 u_2\} du_1 du_2.$$

- An estimator for ρ_S is given by the sample correlation on the ranks of the data.

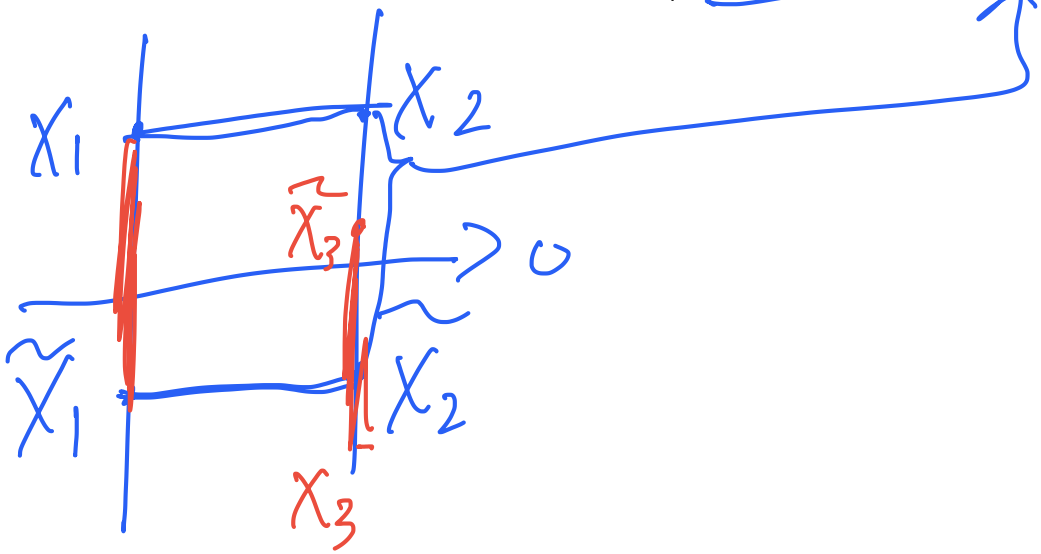
Properties of Spearman's rho

- ρ_S is invariant under any strictly increasing transformations of the random variables.
- $\rho_S(X_1, X_2) = 1 \iff X_1, X_2$ comonotonic.
- $\rho_S(X_1, X_2) = -1 \iff X_1, X_2$ countermonotonic.

Kendall's tau

- Take an independent copy of (X_1, X_2) denoted $(\widetilde{X}_1, \widetilde{X}_2)$.
- The Kendall's tau is defined as the probability of concordance

$$\rho_\tau(X_1, X_2) = 2P\left((X_1 - \widetilde{X}_1)(X_2 - \widetilde{X}_2) > 0\right) - 1 \in [-1, 1]$$



$(X_1 - \widetilde{X}_1)(X_2 - \widetilde{X}_2) > 0$

$(X_1 - \widetilde{X}_1)(X_3 - \widetilde{X}_3) < 0$

> 0 < 0

Kendall's tau

- Take an independent copy of (X_1, X_2) denoted $(\widetilde{X}_1, \widetilde{X}_2)$.
- The Kendall's tau is defined as the probability of concordance

$$\rho_\tau(X_1, X_2) = 2P\left((X_1 - \widetilde{X}_1)(X_2 - \widetilde{X}_2) > 0\right) - 1 \in [-1, 1]$$

- Once can show that this is equivalent to

$$\rho_\tau(X_1, X_2) \quad = \quad 4 \int_0^1 \int_0^1 C(u_1, u_2) dC(u_1, u_2) - 1.$$

- The properties of Kendall's tau are similar to those of Spearman's rho.

Kendall's tau

- Take an independent copy of (X_1, X_2) denoted $(\widetilde{X}_1, \widetilde{X}_2)$.
- The Kendall's tau is defined as the probability of concordance

$$\rho_\tau(X_1, X_2) = 2P\left((X_1 - \widetilde{X}_1)(X_2 - \widetilde{X}_2) > 0\right) - 1 \in [-1, 1]$$

- Once can show that this is equivalent to

$$\rho_\tau(X_1, X_2) \quad = \quad 4\int_0^1 \int_0^1 C(u_1, u_2) dC(u_1, u_2) - 1.$$

- The properties of Kendall's tau are similar to those of Spearman's rho.
- A sample version of ρ_τ is given by

$$\widehat{\rho}_\tau(X_1, X_2) = \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} \text{sign}\left[(X_{i,1} - X_{j,1})(X_{i,2} - X_{j,2})\right].$$

Kendall's tau

- Take an independent copy of (X_1, X_2) denoted $(\widetilde{X}_1, \widetilde{X}_2)$.
- The Kendall's tau is defined as the probability of concordance

$$\rho_\tau(X_1, X_2) = 2P\left((X_1 - \widetilde{X}_1)(X_2 - \widetilde{X}_2) > 0\right) - 1 \in [-1, 1]$$

- Once can show that this is equivalent to

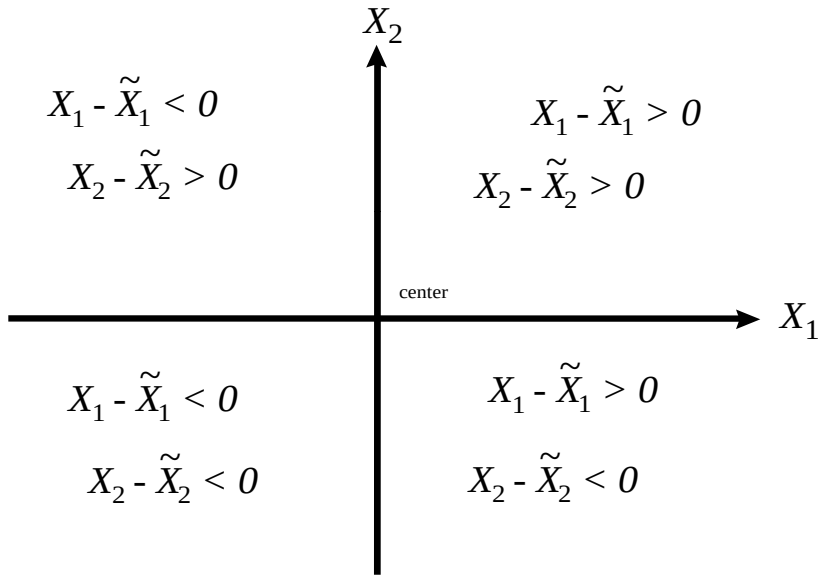
$$\rho_\tau(X_1, X_2) \quad = \quad 4\int_0^1 \int_0^1 C(u_1, u_2) dC(u_1, u_2) - 1.$$

- The properties of Kendall's tau are similar to those of Spearman's rho.
- A sample version of ρ_τ is given by

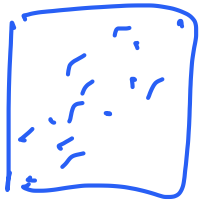
$$\widehat{\rho}_\tau(X_1, X_2) = \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} \text{sign}\left[(X_{i,1} - X_{j,1})(X_{i,2} - X_{j,2})\right].$$

- All dependence measures ρ, ρ_S and ρ_τ can be used to estimate the copula parameters θ_C of a parametric family of copulas by comparing the empirical to the model versions.

Illustration

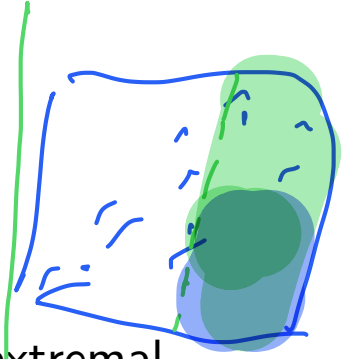


Tail dependence



copulas dependentance

=> now we only want estimate part of this



- Another important measure is given by the tail dependence or extremal dependence (bivariate case).
- When the limit exists, the coefficient of upper tail dependence is

$$\lambda_u(X_1, X_2) = \lim_{q \to 1} P(X_2 > F_2^{-1}(q) \mid X_1 > F_1^{-1}(q)),$$

- Analogously the coefficient of lower tail dependence is

$$\lambda_\ell(X_1, X_2) = \lim_{q \to 0} P\left(X_2 \leq F_2^{-1}(q) \mid X_1 \leq F_1^{-1}(q)\right).$$

q_0

$x_0 > q_0$

if $\lambda_u(X_1, X_2)$'s positive we get always some points in green area.

only look this

$F_1(x_1) > q_0$

but we only need green part so.

Tail dependence

- Another important measure is given by the tail dependence or extremal dependence (bivariate case).
- When the limit exists, the coefficient of upper tail dependence is

$$\lambda_u(X_1, X_2) = \lim_{q \to 1} P(X_2 > F_2^{-1}(q) \mid X_1 > F_1^{-1}(q)),$$

- Analogously the coefficient of lower tail dependence is

$$\lambda_\ell(X_1, X_2) = \lim_{q \to 0} P\left(X_2 \leq F_2^{-1}(q) \mid X_1 \leq F_1^{-1}(q)\right).$$

- When $\lambda_u > 0$ or $\lambda_\ell > 0$ we call (X_1, X_2) upper or lower tail dependent, respectively.
- If $\lambda_u = 0$ or $\lambda_\ell = 0$ we call (X_1, X_2) upper or lower tail independent, respectively.

Copulas and tail dependence

- For a bivariate copula $U \sim C$ let the survival copula be $\overline{C}(u, v) = P(U_1 > u, U_2 > v)$.
- Using the corresponding copulas, we have

$$\begin{aligned} \lambda_u &= \lim_{q \to 1} \frac{\overline{C}(q, q)}{1 - q} = \lim_{q \to 1} \frac{1 - 2q + C(q, q)}{1 - q}, \\ \lambda_\ell &= \lim_{q \to 0} \frac{C(q, q)}{q}. \end{aligned}$$

- Note that $\lambda_u \in [0, 1]$ and $\lambda_\ell \in [0, 1]$.
- Basically, we need to look at the slope of the copula when approaching $(0, 0)$ or $(1, 1)$ along the diagonal.

Example: Clayton copula

- We have

$$C_\beta^{Cl}(q,q) = \left(2q^{-\beta} - 1\right)^{-1/\beta}.$$

$$\begin{aligned}
\lambda_\ell &= \lim_{q\to 0} \frac{C(q,q)}{q} = \lim_{q\to 0} \frac{\left(2q^{-\beta} - 1\right)^{-1/\beta}}{q} \\
&= \lim_{q\to 0} \left(q^\beta\right)^{-1/\beta} \left(2q^{-\beta} - 1\right)^{-1/\beta} \\
&= \lim_{q\to 0} \left(2 - q^\beta\right)^{-1/\beta} = 2^{-1/\beta}
\end{aligned}$$

- The Clayton copula is lower tail dependent for $\beta > 0$.

Further properties

- For elliptical copulas $\lambda_u = \lambda_\ell =: \lambda$.
- The t copula is tail dependent when $\rho > -1$.

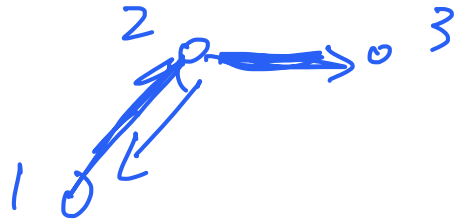
$$\lambda = 2\bar{t}_{\nu+1}\left(\sqrt{\nu+1}\sqrt{1-\rho}/\sqrt{1+\rho}\right).$$

- The Gumbel copula is upper tail dependent for $\beta > 1$.

$$\lambda_u = 2 - 2^{1/\beta}.$$

- The Gaussian copula is upper and lower tail independent, that is, $\lambda_u = \lambda_\ell = 0$.

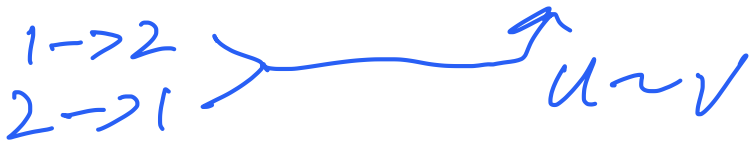
Graphs: notation and terminology



$V = \{1, 2, 3\}$

$E = \{(1,2), (2,3)\}$

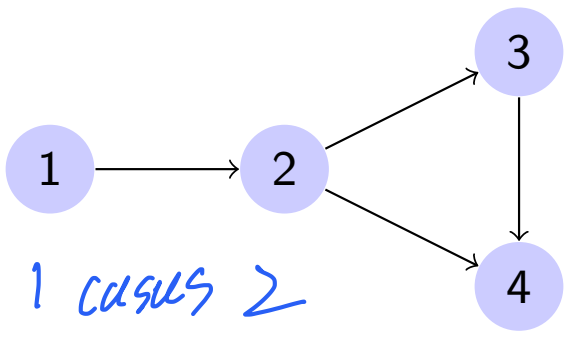
- A graph $G = (V, E)$ is a pair with
  - a set of nodes/vertices $V = \{1, \ldots, p\}$;
  - a set of edges $E \subset V \times V$ between the nodes.
- Edges with $(u, v) \in E$ and $(v, u) \in E$ are called undirected and we write $u \sim v$.



Graphs: notation and terminology

- A graph $G = (V, E)$ is a pair with
  - a set of nodes/vertices $V = \{1, \ldots, p\}$;
  - a set of edges $E \subset V \times V$ between the nodes.
- Edges with $(u, v) \in E$ and $(v, u) \in E$ are called undirected and we write $u \sim v$.
- Edges with $(u, v) \in E$ but $(v, u) \notin E$ are called directed and we write $u \to v$.
- Directed graphs only have directed edges. They are important in the field of causal inference; e.g., Peters et al. (*Elements of Causal Inference*, 2017)

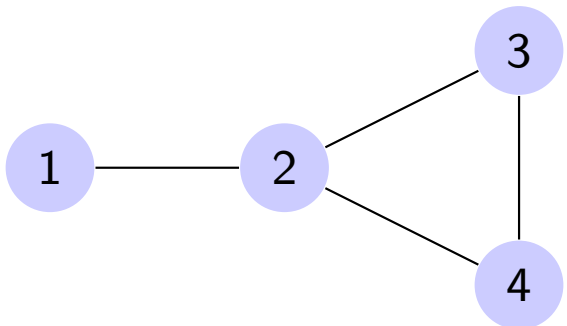
Example for directed graph: $V = \{1, \ldots, 4\}$, $E =$??.



Undirected graphs

- In this course we will only consider **undirected graphs**, that is, all edges are undirected.
- To simplify notation, for $u \sim v$ we write either $(u, v) \in E$ or $(v, u) \in E$, but always assume that both directions are in the edge set.
- Two nodes $u \sim v$ are called adjacent or neighbors.
- See Lauritzen (*Graphical Models*, 1996) for details.

Example for undirected graph: $V = \{1, \ldots, 4\}$, $E =$??.



Terminology

- A subset $A \subset V$ of vertices is called `complete`, if all vertices are adjacent.

Terminology

- A subset $A \subset V$ of vertices is called **complete**, if all vertices are adjacent.
- A complete subset that is maximal is called a **clique**.

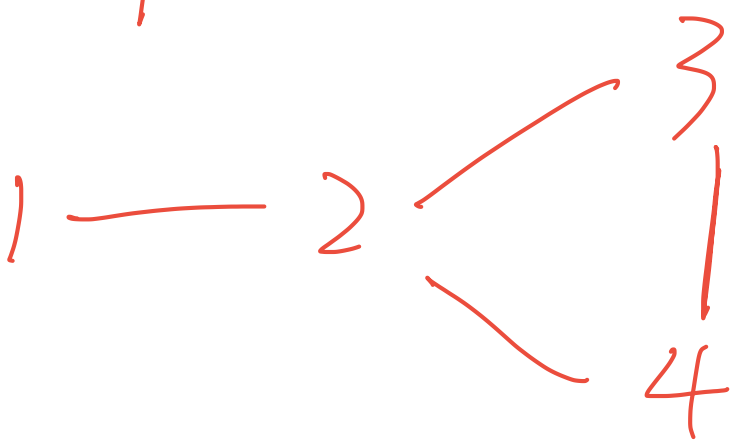
Terminology

- A subset $A \subset V$ of vertices is called **complete**, if all vertices are adjacent.
- A complete subset that is maximal is called a **clique**.
- A **path** from node u to node v of length n is a sequence $u = w_0, w_1 \ldots, w_n = v$, where $w_1, \ldots, w_n$ are distinct nodes and $(w_{i-1}, w_i) \in E$ for all $i = 1, \ldots, n$.

path from 3 to 1

{ length 2 w_3 w_2 w_1

length 3 w_3 w_4 w_2 w_1



Terminology

- A subset $A \subset V$ of vertices is called `complete`, if all vertices are adjacent.
- A complete subset that is maximal is called a `clique`.
- A `path` from node u to node v of length n is a sequence $u = w_0, w_1 \ldots, w_n = v$, where $w_1, \ldots, w_n$ are distinct nodes and $(w_{i-1}, w_i) \in E$ for all $i = 1, \ldots, n$.
- Two nodes are connected if there exists a path between them.

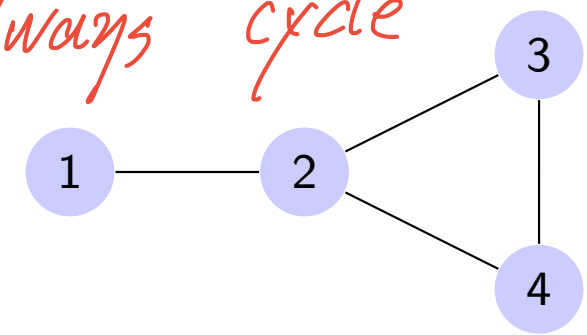
Terminology

- A subset $A \subset V$ of vertices is called **complete**, if all vertices are adjacent.
- A complete subset that is maximal is called a **clique**.
- A **path** from node u to node v of length n is a sequence $u = w_0, w_1 \ldots, w_n = v$, where $w_1, \ldots, w_n$ are distinct nodes and $(w_{i-1}, w_i) \in E$ for all $i = 1, \ldots, n$.
- Two nodes are connected if there exists a path between them.
- A cycle is a path with $u = v$, i.e., a path that begins and ends in the same point.

3243 ≠ 2342

Example: Find the complete subsets, cliques and cycles.

clique is not always cycle

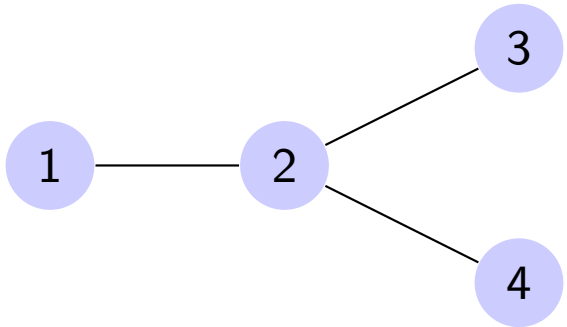


2 3 4

Examples for graphs

- A **tree** is a connected, undirected graph without cycles, or equivalently, there exists a unique path between any two nodes.

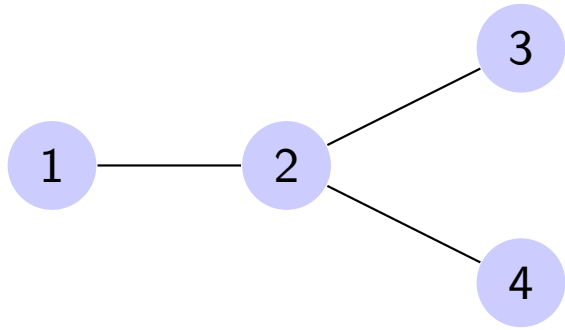
Example for a tree



Examples for graphs

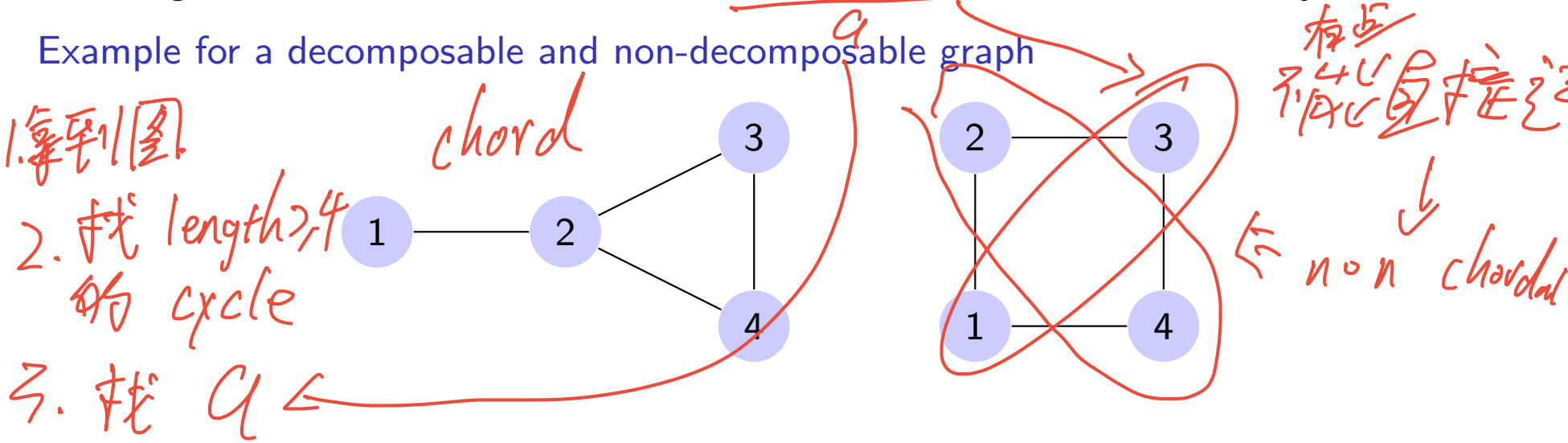
- A **tree** is a connected, undirected graph without cycles, or equivalently, there exists a unique path between any two nodes.

Example for a tree



- A **decomposable/chordal/triangulated** graph is a graph where each cycle of length $n \geq 4$ has a chord, i.e., two non-consecutive nodes that are adjacent.

Example for a decomposable and non-decomposable graph

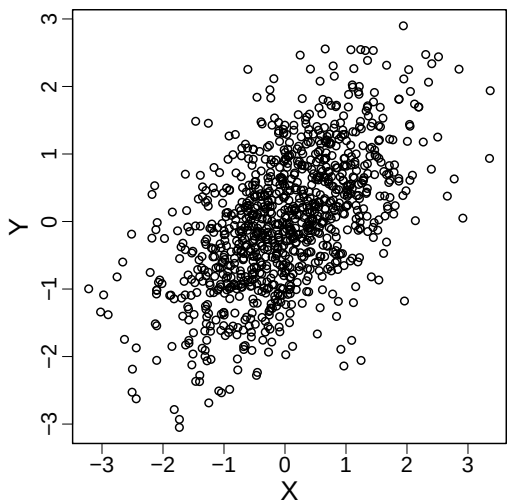


Conditional independence

- Recall: Two random vectors with joint density $f(x, y)$ are independent of $f(x, y) = f(x)f(y)$.
- This is a very strong assumption, and many variables cannot be assumed to be independent.

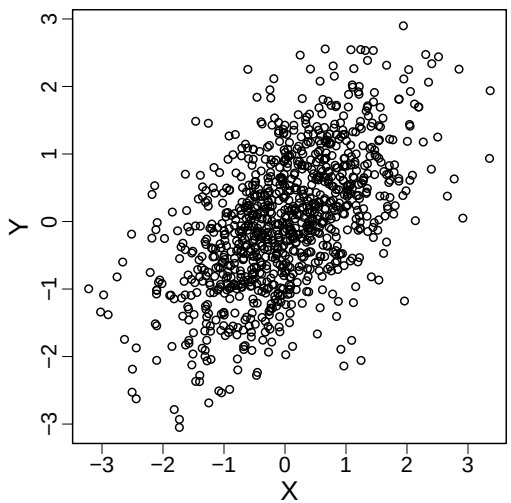
Conditional independence

- Recall: Two random vectors with joint density $f(x, y)$ are independent of $f(x, y) = f(x)f(y)$.
- This is a very strong assumption, and many variables cannot be assumed to be independent.
- Example 1: $X =$ "Ice cream sales" and $Y =$ "Violent crime"



Conditional independence

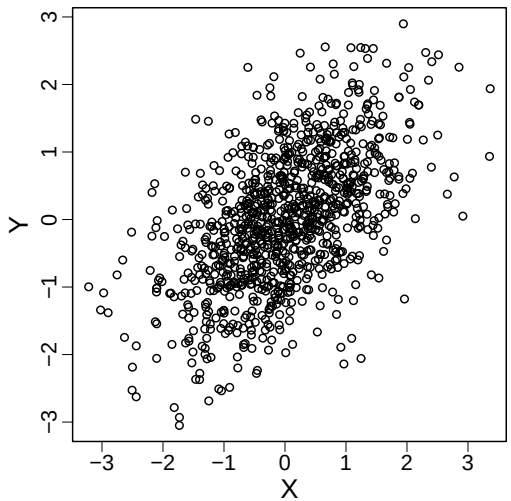
- Recall: Two random vectors with joint density $f(x, y)$ are independent of $f(x, y) = f(x)f(y)$.
- This is a very strong assumption, and many variables cannot be assumed to be independent.
- Example 1: $X =$ "Ice cream sales" and $Y =$ "Violent crime"
- Example 2: $X =$ "Student 1 late" and $Y =$ "Student 2 late"

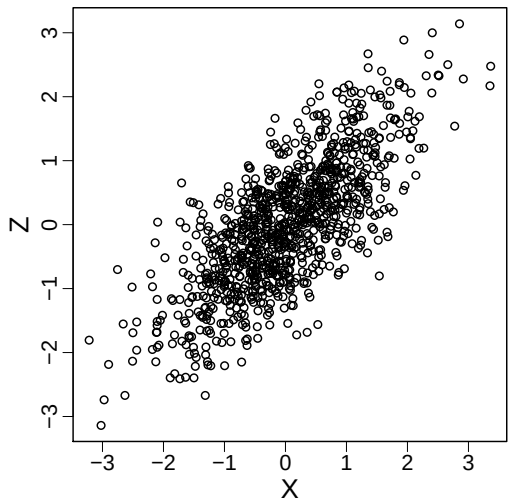


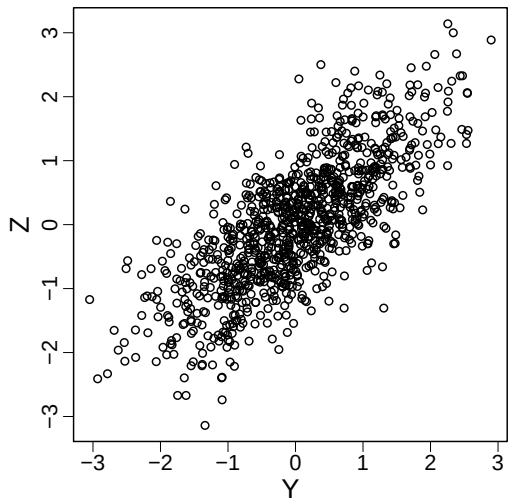
Conditional independence

- Recall: Two random vectors with joint density $f(x, y)$ are independent of $f(x, y) = f(x)f(y)$.
- This is a very strong assumption, and many variables cannot be assumed to be independent.
- Example 1: $X =$ "Ice cream sales" and $Y =$ "Violent crime"
- Example 2: $X =$ "Student 1 late" and $Y =$ "Student 2 late"
- Conditional independence of X and Y given a third variable Z means:

"If I already know Z, then information on Y is irrelevant for X".







Conditional independence

- Let X, Y and Z be random vectors with joint density $f_{X,Y,Z}$. We say that X is **conditionally independent** of Y given Z, if

$$f_{X \mid Z,Y}(x \mid z, y) = f_{X|Z}(x \mid z).$$

  In this case we write $X \perp\!\!\!\perp Y \mid Z$.

Conditional independence

- Let X, Y and Z be random vectors with joint density $f_{X,Y,Z}$. We say that X is **conditionally independent** of Y given Z, if

$$f_{X|Y,Z}(x \mid z, y) = f_{X|Z}(x \mid z).$$

  In this case we write $X \perp\!\!\!\perp Y \mid Z$.

- There are many equivalent definitions, e.g.,

$$f_{X,Y|Z}(x, y \mid z) = f_{X|Z}(x \mid z) f_{Y|Z}(y \mid z).$$

  or

$$f_{X,Y,Z}(x, y, z) = h(x, z) k(y, z),$$

  for suitable functions h and k.

$$f(x|z,y) = \frac{f(z,x,y)}{f(z,y)} = \frac{f(x,y|z)f(z)}{f(z,y)} = \frac{f(x|z)f(y|z)f(z)}{f(y,z)} = \frac{f(y,z)}{f(z)} \quad = f(x|z)$$

conditional density

Conditional independence

- Let X, Y and Z be random vectors with joint density $f_{X,Y,Z}$. We say that X is **conditionally independent** of Y given Z, if

$$f_{X,Y|Z}(x \mid z, y) = f_{X|Z}(x \mid z).$$

  In this case we write $X \perp\!\!\!\perp Y \mid Z$.

- There are many equivalent definitions, e.g.,

$$f_{X,Y|Z}(x, y \mid z) = f_{X|Z}(x \mid z) f_{Y|Z}(y \mid z).$$

  or

$$f_{X,Y,Z}(x, y, z) = h(x, z) k(y, z),$$

  for suitable functions h and k.

- This clearly generalizes the classical notion of independence, in which case we would have that Z is a trivial, empty vector.

The pairwise Markov property

- How do we link conditional independence to graph structures?
- We consider an undirected graph $G = (V, E)$ and a collection of random variables $X = (X_u)_{u \in V}$ indexed by the nodes of the graph.

The pairwise Markov property

- How do we link conditional independence to graph structures?
- We consider an undirected graph $G = (V, E)$ and a collection of random variables $X = (X_u)_{u \in V}$ indexed by the nodes of the graph.
- We say that X satisfies the **pairwise Markov property** if for any $u, v \in V$ with $(u, v) \notin E$ it holds that

$$X_u \perp\!\!\!\perp X_v \mid X_{V \setminus \{u,v\}}.$$

- In this case we also say that X is a graphical model on G.

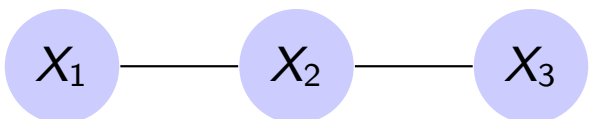
The pairwise Markov property

- How do we link conditional independence to graph structures?
- We consider an undirected graph $G = (V, E)$ and a collection of random variables $X = (X_u)_{u \in V}$ indexed by the nodes of the graph.
- We say that X satisfies the **pairwise Markov property** if for any $u, v \in V$ with $(u, v) \notin E$ it holds that

$$X_u \perp\!\!\!\perp X_v \mid X_{V \setminus \{u,v\}}.$$

- In this case we also say that X is a graphical model on G.

Example:



$$X_1 \perp\!\!\!\perp X_3 \mid X_2$$

$$f(x_1, x_2, x_3) = f(x_1 \mid x_2, x_3) f(x_2, x_3) = f(x_1 \mid x_2) f(x_2 \mid x_3) f(x_3) = f(x_1, x_2) f(x_2, x_3) / f(x_2)$$

The global Markov property

- Pairwise Markov property: for any $u, v \in V$ with $(u, v) \notin E$ it holds that

$$X_u \perp\!\!\!\perp X_v \mid X_{V \setminus \{u,v\}}.$$

The global Markov property

- Pairwise Markov property: for any $u, v \in V$ with $(u, v) \notin E$ it holds that

$$X_u \perp\!\!\!\perp X_v \mid X_{V \setminus \{u,v\}}.$$

- For disjoint subsets $A, B, S \subset V$, S is said to separate A and B if all paths from a node $u \in A$ and $v \in B$ intersect S.

The global Markov property

- Pairwise Markov property: for any $u, v \in V$ with $(u, v) \notin E$ it holds that

$$X_u \perp\!\!\!\perp X_v \mid X_{V \setminus \{u,v\}}.$$

- For disjoint subsets $A, B, S \subset V$, S is said to separate A and B if all paths from a node $u \in A$ and $v \in B$ intersect S.
- We say that X satisfies the **global Markov property** on G if for any disjoint subsets $A, B, S \subset V$ such that S separates A and B in G, we have

$$X_A \perp\!\!\!\perp X_B \mid X_S.$$

The global Markov property

- Pairwise Markov property: for any $u, v \in V$ with $(u, v) \notin E$ it holds that

$$X_u \perp\!\!\!\perp X_v \mid X_{V \setminus \{u,v\}}.$$

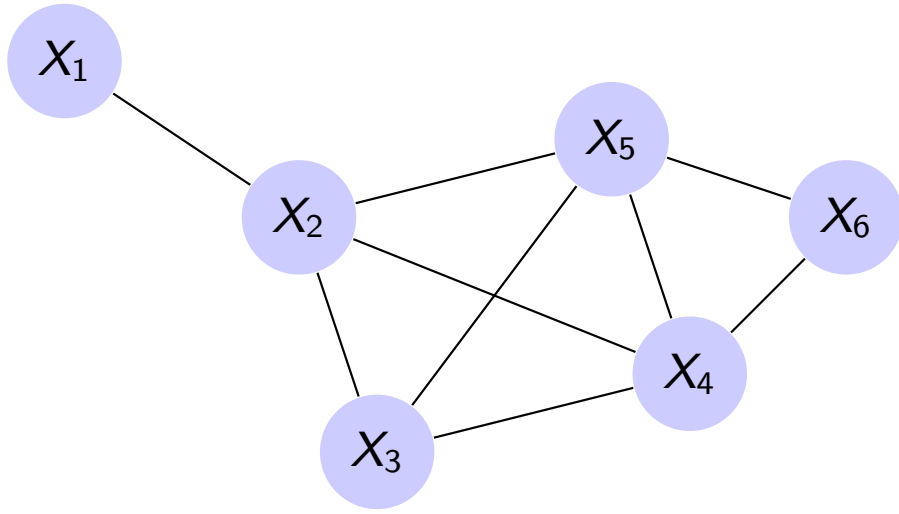
- For disjoint subsets $A, B, S \subset V$, S is said to separate A and B if all paths from a node $u \in A$ and $v \in B$ intersect S.
- We say that X satisfies the **global Markov property** on G if for any disjoint subsets $A, B, S \subset V$ such that S separates A and B in G, we have

$$X_A \perp\!\!\!\perp X_B \mid X_S.$$

- It turns out, that if $X = (X_u)_{u \in V}$ has a positive and continuous density, then pairwise and global Markov properties are **equivalent** .

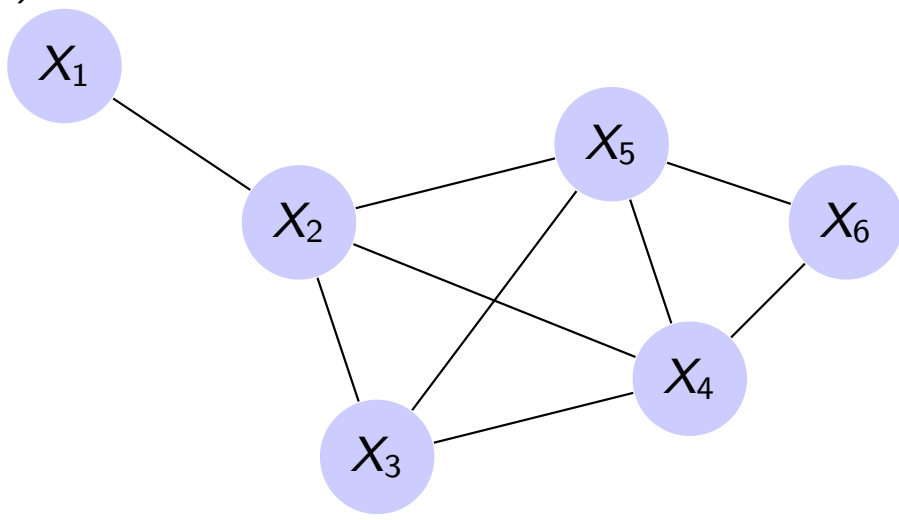
Examples

- List all conditional independencies and simplify the joint density $f(x_1, x_2, x_3, x_4, x_5, x_6)$:



Examples

- List all conditional independencies and simplify the joint density $f(x_1, x_2, x_3, x_4, x_5, x_6)$:



- Draw the corresponding graph:

$$f(x_1, x_2, x_3, x_4, x_5) = f(x_3)f(x_5 \mid x_3)f(x_1, x_4 \mid x_5)f(x_2)$$

The Hammersley–Clifford theorem

Theorem

Let $X = (X_u)_{u \in V}$ be a random vector with positive and continuous density f on $\mathcal{X}$. Then the vector X satisfies the **pairwise Markov property** on the graph $G = (V, E)$ if and only if its **density factorizes** as

$$f(x) = \prod_{C \in \mathcal{C}} \psi_C(x_C), \qquad x \in \mathcal{X},$$

for suitable functions ψ_C with arguments $x_C = (x_u)_{u \in C}$, where $\mathcal{C}$ is the set of all cliques of G.

The Hammersley–Clifford theorem

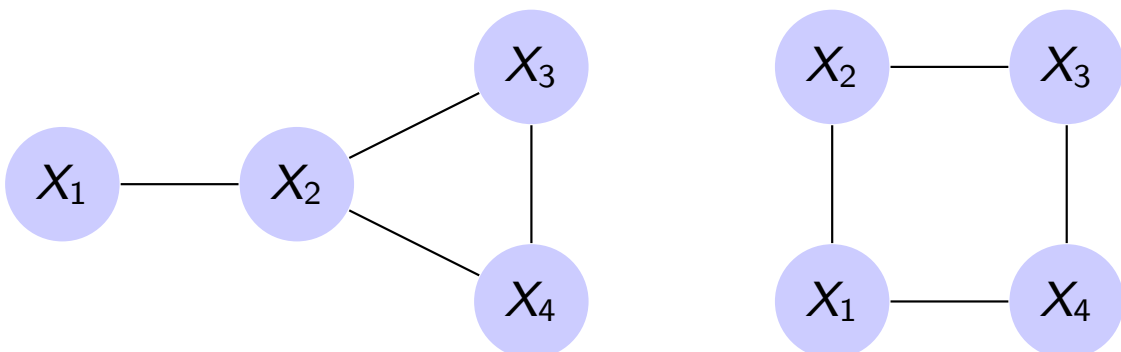
Theorem

Let $X = (X_u)_{u \in V}$ be a random vector with positive and continuous density f on $\mathcal{X}$. Then the vector X satisfies the **pairwise Markov property** on the graph $G = (V, E)$ if and only if its **density factorizes** as

$$f(x) = \prod_{C \in \mathcal{C}} \psi_C(x_C), \qquad x \in \mathcal{X},$$

for suitable functions ψ_C with arguments $x_C = (x_u)_{u \in C}$, where $\mathcal{C}$ is the set of all cliques of G.

Examples:



Hammersley–Clifford for decomposable graphs

Corollary

Let $X = (X_u)_{u \in V}$ be a random vector with positive and continuous density f on $\mathcal{X}$. If X satisfies the pairwise Markov property on the **decomposable** graph $G = (V, E)$ then

$$f(x) = \frac{\prod_{C \in \mathcal{C}} f_C(x_C)}{\prod_{D \in \mathcal{D}} f_D(x_D)}, \qquad x \in \mathcal{X},$$

where $\mathcal{D}$ is a multiset containing intersections between the maximal cliques called separator sets.

Hammersley–Clifford for decomposable graphs

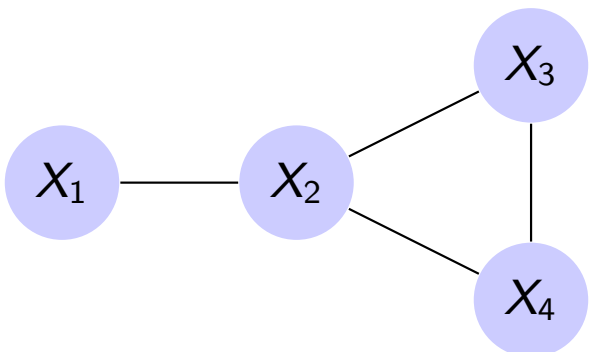
Corollary

Let $X = (X_u)_{u \in V}$ be a random vector with positive and continuous density f on $\mathcal{X}$. If X satisfies the pairwise Markov property on the **decomposable** graph $G = (V, E)$ then

$$f(x) = \frac{\prod_{C \in \mathcal{C}} f_C(x_C)}{\prod_{D \in \mathcal{D}} f_D(x_D)}, \qquad x \in \mathcal{X},$$

where $\mathcal{D}$ is a multiset containing intersections between the maximal cliques called separator sets.

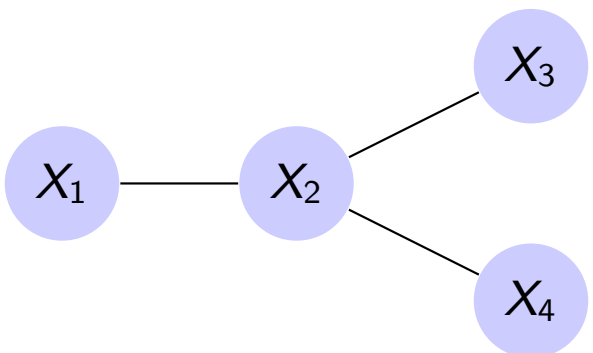
Example:



Tree graphical models

- If X is a graphical model on the tree $T = (V, E)$, then its density has the simple form

$$f(x) = \prod_{(u,v)\in E} \frac{f_{uv}(x_u, x_v)}{f_u(x_u) f_v(x_v)} \prod_{u\in V} f_u(x_u), \qquad x \in \mathcal{X}.$$



Recap: formulas for Gaussian distributions

- If $X \sim N_p(\mu, \Sigma)$ can be partitioned into $(X_{(1)}, X_{(2)}) \in \mathbb{R}^{p_1+p_2}$, $\mu = (\mu_{(1)}, \mu_{(2)})$ and

$$\Sigma = \begin{bmatrix} \Sigma_{(11)} & \Sigma_{(12)} \\ \Sigma_{(21)} & \Sigma_{(22)} \end{bmatrix},$$

  then the conditional distribution of $X_{(2)}$ given $X_{(1)} = x_{(1)}$ is the density of p_2-variate normal distribution

$$N_{p_2}\left(\mu_{(2)} + \Sigma_{(21)}\Sigma_{(11)}^{-1}(x_1 - \mu_{(1)}), \Sigma_{(22)} - \Sigma_{(21)}\Sigma_{(11)}^{-1}\Sigma_{(12)}\right)$$

Gaussian graphical models

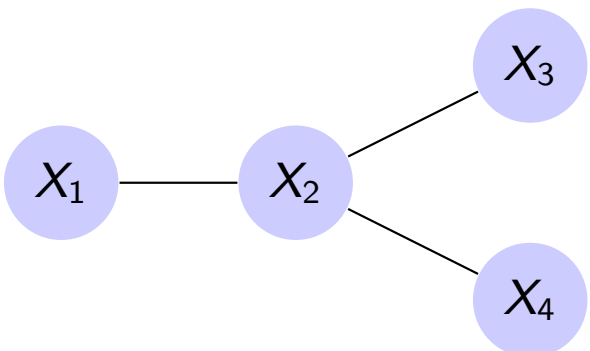
Proposition

Let Σ be a positive-definite $p \times p$-matrix and $\mu \in \mathbb{R}^p$, and $X \sim N_p(\mu, \Sigma)$. For any $i, j \in \{1, \ldots, p\}$ with $i \neq j$ we have

$$X_i \perp\!\!\!\perp X_j \mid X_{\setminus\{i,j\}} \Leftrightarrow K_{ij} = 0,$$

where $K = \{K_{ij}\}_{i,j=1,\ldots p} = \Sigma^{-1}$ is called the concentration/precision matrix.

Example:



$$K = \begin{bmatrix} * & * & 0 & 0 \\ * & * & * & * \\ 0 & * & * & 0 \\ 0 & * & 0 & * \end{bmatrix}$$

Structure learning

- In most applications, the conditional independencies and the underlying graph structure are `not known`.

Structure learning

- In most applications, the conditional independencies and the underlying graph structure are `not known`.
- Therefore, the best graphical structure should be learned from observations.

Maximum likelihood trees

Recall the density for a tree $T = (V, E)$:

$$f(x) = \prod_{(u,v)\in E} \frac{f_{uv}(x_u, x_v)}{f_u(x_u) f_v(x_v)} \prod_{u \in V} f_u(x_u), \qquad x \in \mathcal{X}.$$

- Suppose we have observations $X_1, \ldots, X_n \sim N_p(\mu, \Sigma)$ and it is known that the concentration matrix $K = \Sigma^{-1}$ has a **tree structure**, but the specific tree structure T and the parameters are unknown.

Maximum likelihood trees

Recall the density for a tree $T = (V, E)$:

$$f(x) = \prod_{(u,v)\in E} \frac{f_{uv}(x_u, x_v)}{f_u(x_u) f_v(x_v)} \prod_{u \in V} f_u(x_u), \qquad x \in \mathcal{X}.$$

- Suppose we have observations $X_1, \ldots, X_n \sim N_p(\mu, \Sigma)$ and it is known that the concentration matrix $K = \Sigma^{-1}$ has a **tree structure**, but the specific tree structure T and the parameters are unknown.
- For a given tree T, let $\mathcal{S}^+(T)$ denote the space of all positive-definite concentration matrices that factorize according to the tree T.

Maximum likelihood trees

Recall the density for a tree $T = (V, E)$:

$$f(x) = \prod_{(u,v)\in E} \frac{f_{uv}(x_u, x_v)}{f_u(x_u) f_v(x_v)} \prod_{u\in V} f_u(x_u), \qquad x \in \mathcal{X}.$$

- Suppose we have observations $X_1, \ldots, X_n \sim N_p(\mu, \Sigma)$ and it is known that the concentration matrix $K = \Sigma^{-1}$ has a **tree structure**, but the specific tree structure T and the parameters are unknown.
- For a given tree T, let $\mathcal{S}^+(T)$ denote the space of all positive-definite concentration matrices that factorize according to the tree T.
- For the fixed tree T, we first maximize the likelihood over all parameters of the concentration matrix for this given tree:

$$\widehat{\ell}(T) = \widehat{\ell}(T \mid X_1, \ldots, X_n) = \max_{K\in\mathcal{S}^+(T)} \widehat{\ell}(K \mid X_1, \ldots, X_n)$$

Maximum likelihood trees

- Thus, the maximized likelihood is

$$\widehat{\ell}(T) = \prod_{(u,v)\in E} \frac{\widehat{f}_{uv}(x_u, x_v)}{\widehat{f}_u(x_u)\widehat{f}_v(x_v)} \prod_{u\in V} \widehat{f}_u(x_u),$$

  where $\widehat{f}_A$ denotes the maximized likelihood based only on data X_{ij}, $j \in A \subset V$, $i = 1, \ldots, n$.

Maximum likelihood trees

- Thus, the maximized likelihood is

$$\widehat{\ell}(T) = \prod_{(u,v)\in E} \frac{\widehat{f}_{uv}(x_u, x_v)}{\widehat{f}_u(x_u)\widehat{f}_v(x_v)} \prod_{u\in V} \widehat{f}_u(x_u),$$

  where $\widehat{f}_A$ denotes the maximized likelihood based only on data X_{ij}, $j \in A \subset V$, $i = 1, \ldots, n$.

- We compute that:

$$\log \frac{\widehat{f}_{uv}(x_u, x_v)}{\widehat{f}_u(x_u)\widehat{f}_v(x_v)} = -\frac{n}{2} \log(1 - r_{uv}^2),$$

  where r_{uv} is the `empirical correlation coefficient`

$$r_{uv} = w_{uv} / \sqrt{w_{uu} w_{vv}},$$

  and $\mathbf{W} = \{w_{uv}, u, v \in V\} = (n-1)\mathbf{S}$ is the Wishart matrix of sums of products.

Maximum likelihood trees

- Define the **empirical correlation weight** w_{uv} of the edge $(u, v) \in E$ as

$$w_{uv} = -\frac{n}{2} \log(1 - r_{uv}^2),$$

and let the total empirical weight of the tree T be

$$w(T) = \sum_{(u,v) \in E} w_{uv}.$$

Maximum likelihood trees

- Define the **empirical correlation weight** w_{uv} of the edge $(u, v) \in E$ as

$$w_{uv} = -\frac{n}{2} \log(1 - r_{uv}^2),$$

and let the total empirical weight of the tree T be

$$w(T) = \sum_{(u,v) \in E} w_{uv}.$$

- The maximized likelihood of all nodes independent, i.e., with no edges, is

$$\widehat{\ell}(\emptyset) = \prod_{u \in V} \widehat{f}_u(x_u).$$

Maximum likelihood trees

- Define the **empirical correlation weight** w_{uv} of the edge $(u, v) \in E$ as

$$w_{uv} = -\frac{n}{2} \log(1 - r_{uv}^2),$$

and let the total empirical weight of the tree T be

$$w(T) = \sum_{(u,v)\in E} w_{uv}.$$

- The maximized likelihood of all nodes independent, i.e., with no edges, is

$$\widehat{\ell}(\emptyset) = \prod_{u\in V} \widehat{f}_u(x_u).$$

- With this notation we have

$$\log \widehat{\ell}(T) - \log \widehat{\ell}(\emptyset) = w(T) = \sum_{(u,v)\in E} w_{uv}.$$

Maximum likelihood trees

- The **maximum likelihood tree** is therefore given by the maximizer

$$\widehat{T} = \operatorname*{arg\,max}_{T=(V,E)} \log \widehat{\ell}(T) - \log \widehat{\ell}(\emptyset) = \operatorname*{arg\,max}_{T=(V,E)} \sum_{(u,v)\in E} w_{uv}.$$

Maximum spanning trees

Maximum spanning tree

Suppose that a positive real number $w_{uv} > 0$ is attached to each pair of nodes (u, v), $u, v \in E$, with $w_{uu} = 0$ and $w_{uv} = w_{vu}$. The **maximum spanning tree** is the tree $T_{\max} = (V, E_{\max})$, $E \subset V \times V$ that maximizes the sum of weights on that tree, i.e.,

$$T_{\max} = \arg\max_{T=(V,E)} \sum_{(u,v)\in E} w_{uv}. \tag{1}$$

Maximum spanning trees

Maximum spanning tree

Suppose that a positive real number $w_{uv} > 0$ is attached to each pair of nodes (u, v), $u, v \in E$, with $w_{uu} = 0$ and $w_{uv} = w_{vu}$. The **maximum spanning tree** is the tree $T_{\max} = (V, E_{\max})$, $E \subset V \times V$ that maximizes the sum of weights on that tree, i.e.,

$$T_{\max} = \arg\max_{T=(V,E)} \sum_{(u,v)\in E} w_{uv}. \tag{1}$$

- Given a set of weights $(w_{uv})_{u,v\in V}$, there are very efficient greedy algorithms to solve (1).

Gaussian trees

Theorem

Let $X \sim N_p(\mu, \Sigma)$ and suppose it satisfies the pairwise Markov property on the tree $T = (V, E)$. Let

$$w_{uv} = -\frac{n}{2} \log(1 - r_{uv}^2),$$

based on observations $X_1, \ldots, X_n \sim N_p(\mu, \Sigma)$, then the maximum spanning tree $T_{\max}$ with these weights is equal to the maximum likelihood tree and it satisfies

$$\mathbb{P}(T_{\max} = T) \to 1, \qquad n \to \infty,$$

that is, the true tree structure is recovered almost surely.

Gaussian trees

Theorem

Let $X \sim N_p(\mu, \Sigma)$ and suppose it satisfies the pairwise Markov property on the tree $T = (V, E)$. Let

$$w_{uv} = -\frac{n}{2} \log(1 - r_{uv}^2),$$

based on observations $X_1, \ldots, X_n \sim N_p(\mu, \Sigma)$, then the maximum spanning tree $T_{\max}$ with these weights is equal to the maximum likelihood tree and it satisfies

$$\mathbb{P}(T_{\max} = T) \to 1, \qquad n \to \infty,$$

that is, the true tree structure is recovered almost surely. Moreover, it follows that

$$T = \underset{T=(V,E)}{\arg\max} \sum_{(u,v)\in E} -\frac{n}{2} \log(1 - \rho_{uv}^2),$$

where ρ_{uv} is the **correlation coefficient** between X_u and X_v, $u, v \in V$.

Kruskal's algorithm

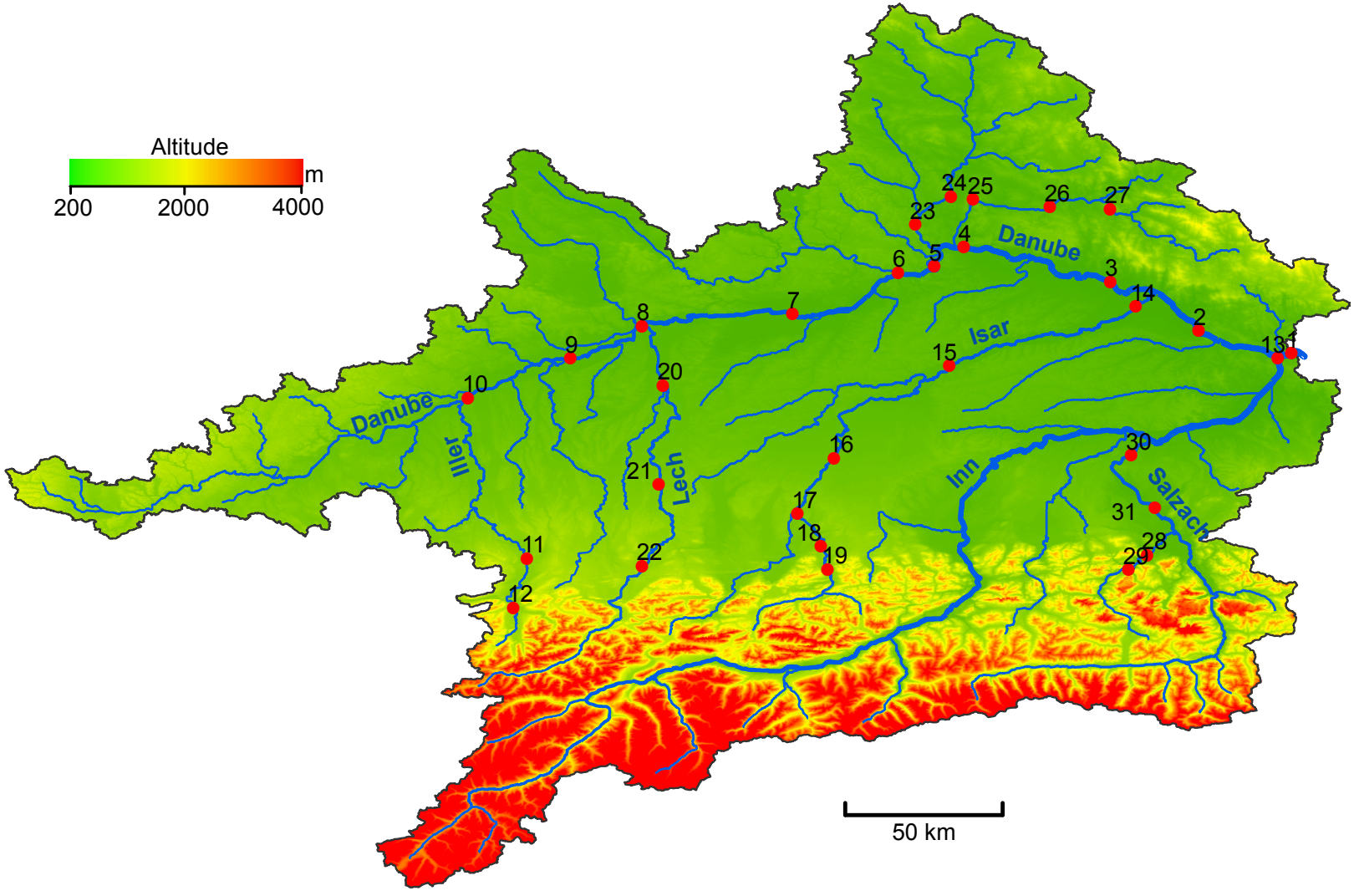
Kruskal's algorithm

Under the notation from above perform the following steps to obtain the maximum spanning tree $T_{\max}$.

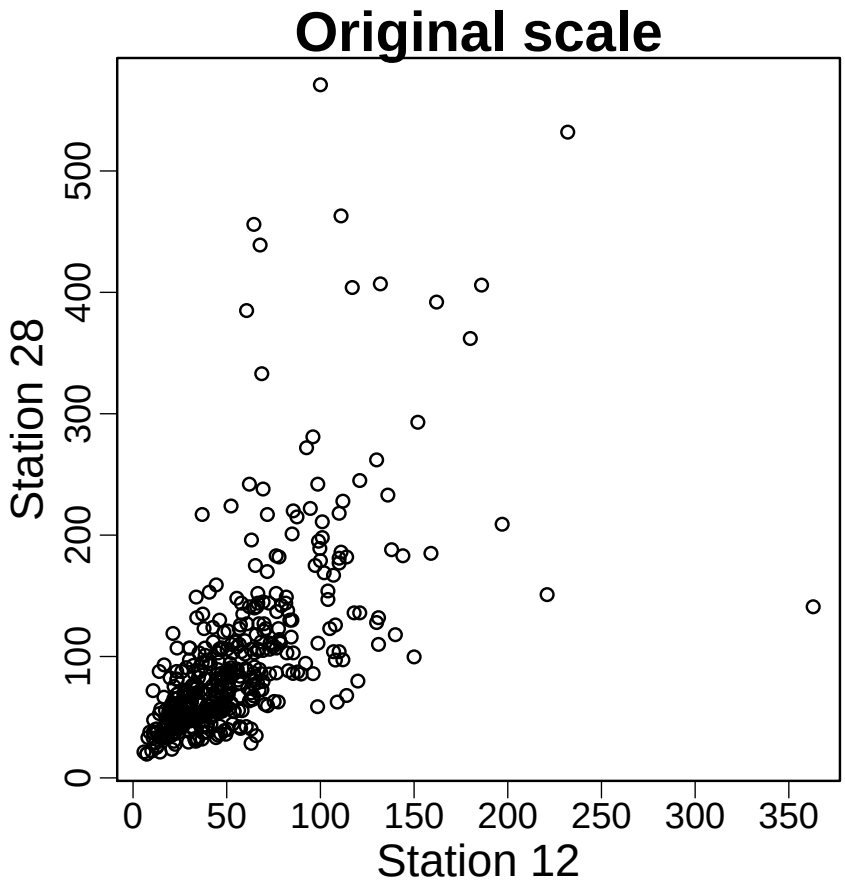
1. Initiate $E_{\max} = \emptyset$.
2. Among all edges in (u, v) not yet chosen, choose the one with the shortest length that does not form any loops with the edges already chosen, and add it to $E_{\max}$.
3. Repeat 2. until no more edges can be added to $E_{\max}$ without creating a loop.

River discharge data

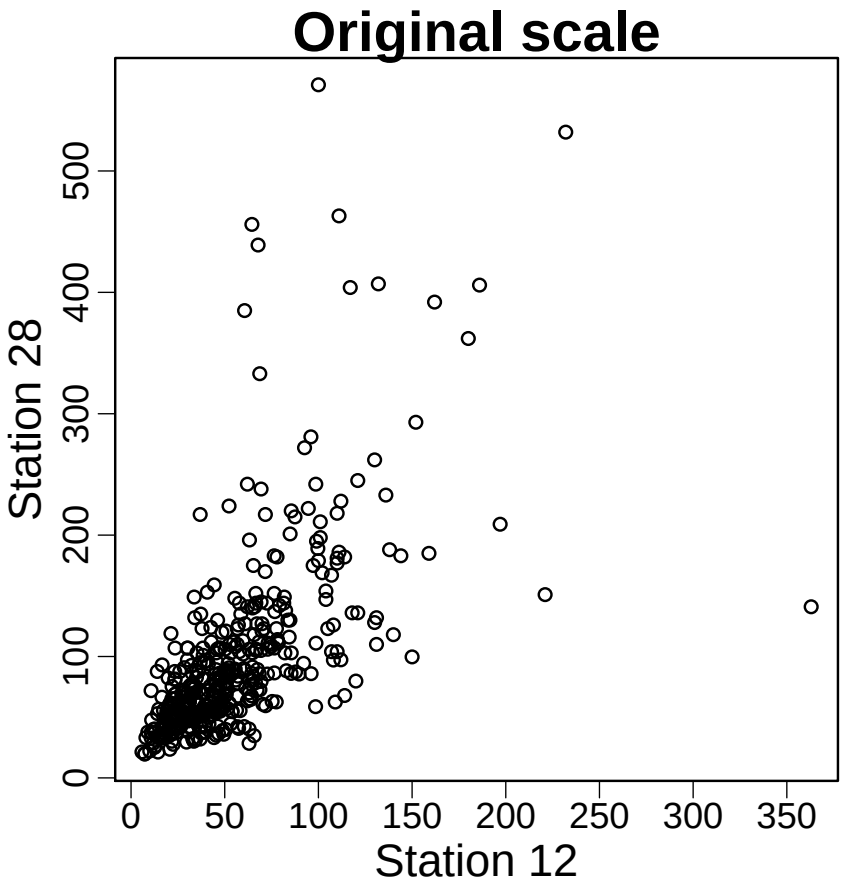
We have $n = 428$ independent observations of $X \in \mathbb{R}^p$, where $p = 31$ is the number of stations on the river network.

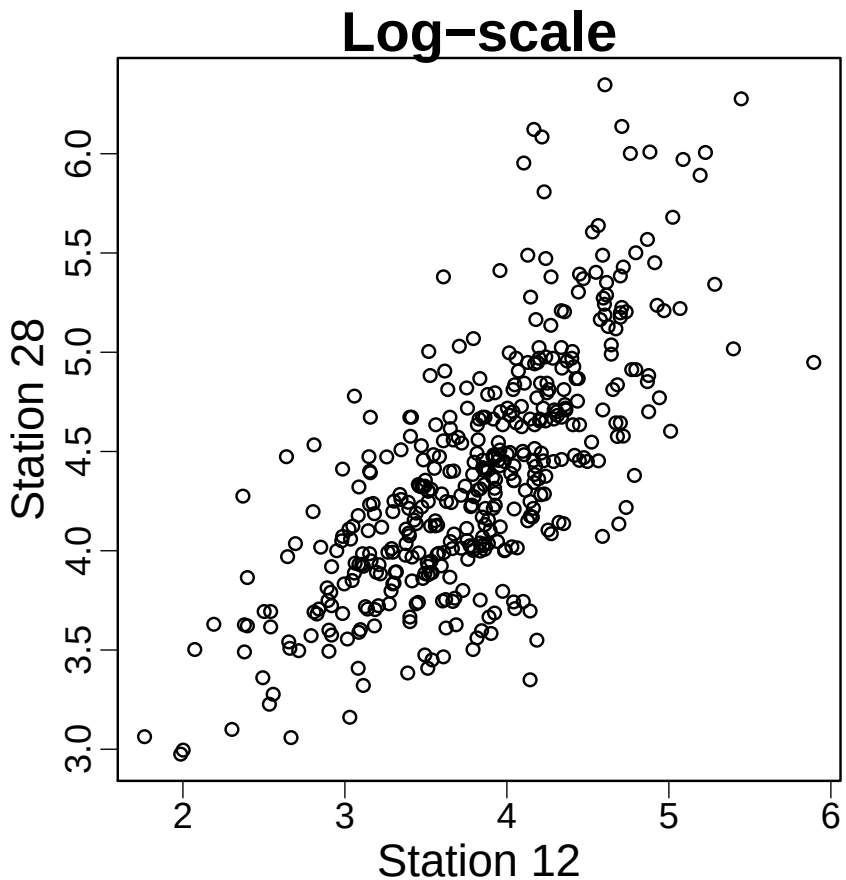


River discharge data

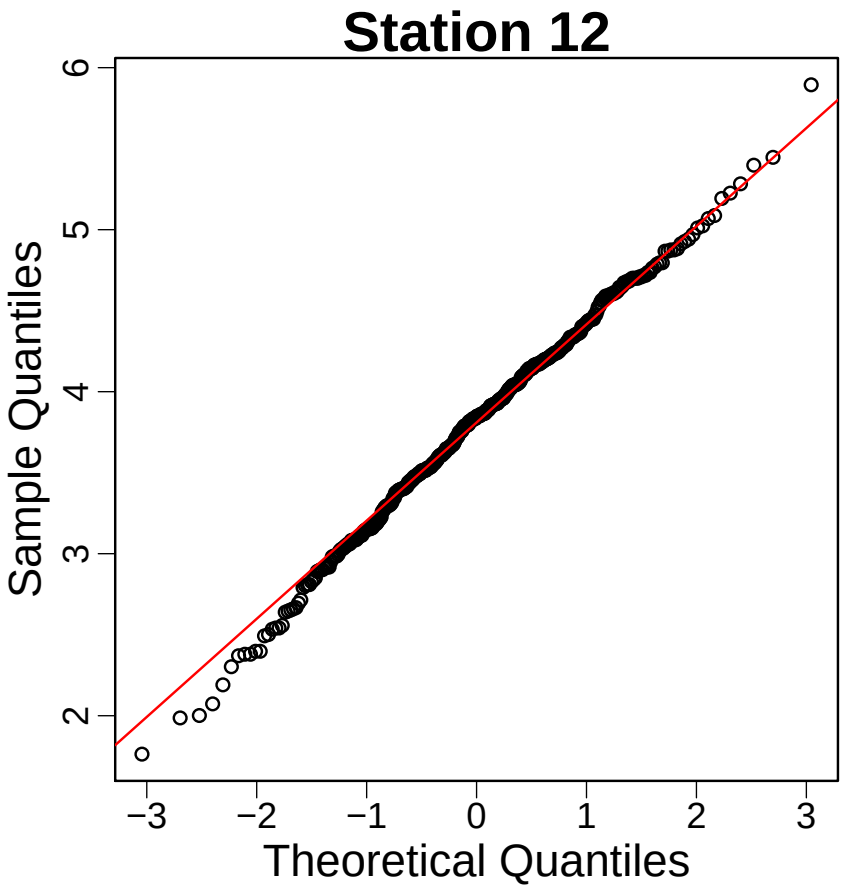


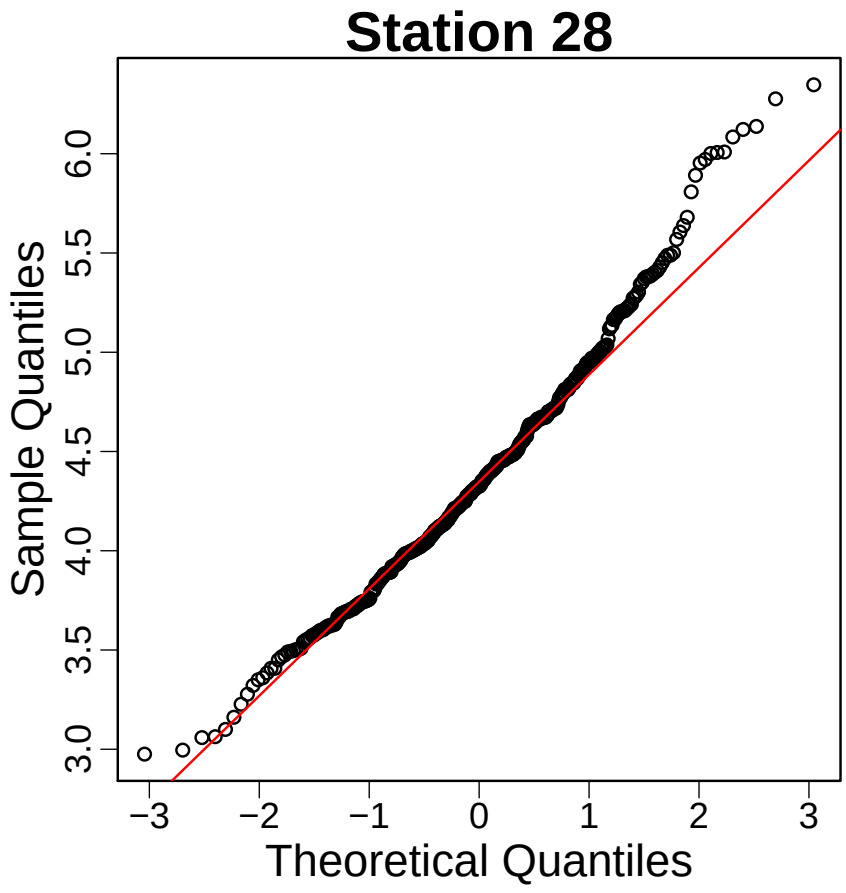
River discharge data



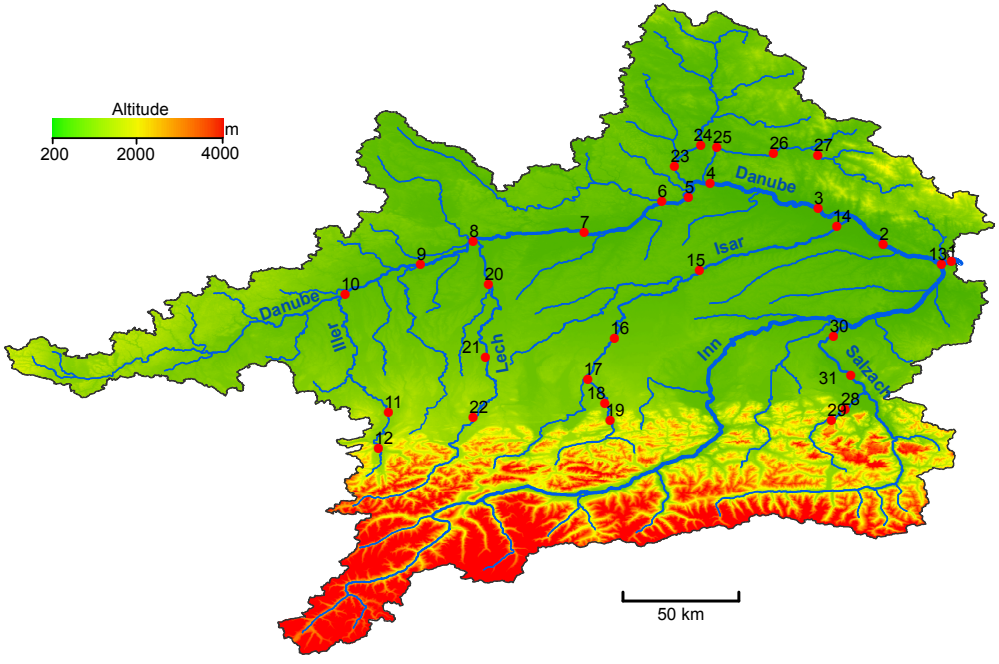


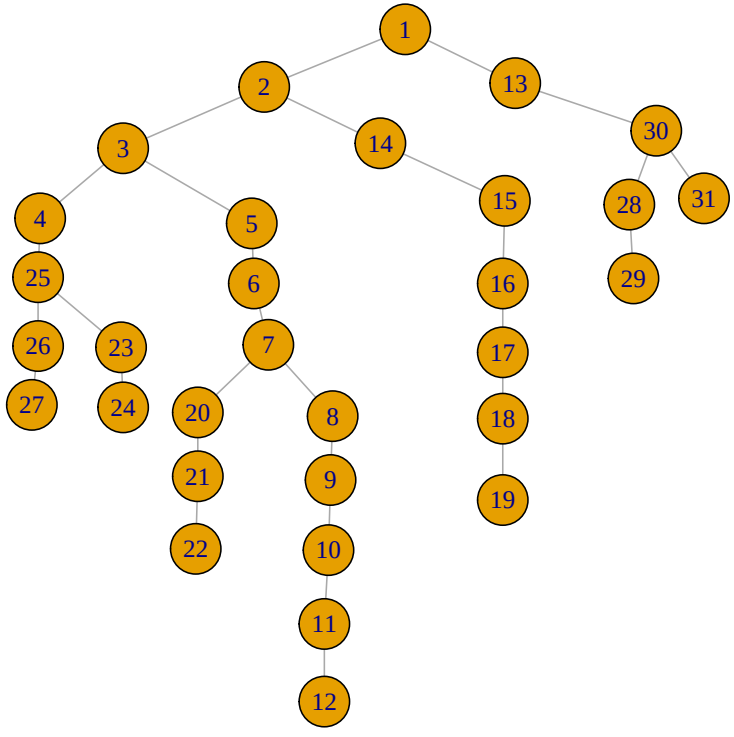
River discharge data





River discharge data: maximum spanning tree





Structure learning for general Gaussian graphical models

Recall the Gaussian log-density for data $X_1, \ldots X_n \sim N_p(\mu, \Sigma)$:

$$\ell(\mu, \Sigma) \propto -\frac{n}{2} \log \det(\Sigma) - \frac{n}{2} \operatorname{tr}(\widehat{\Sigma}\Sigma^{-1}) - \frac{n}{2}(\overline{X} - \mu)^\top \Sigma^{-1}(\overline{X} - \mu).$$

Structure learning for general Gaussian graphical models

Recall the Gaussian log-density for data $X_1, \ldots X_n \sim N_p(\mu, \Sigma)$:

$$\ell(\mu, \Sigma) \propto -\frac{n}{2} \log \det(\Sigma) - \frac{n}{2} \operatorname{tr}(\widehat{\Sigma}\Sigma^{-1}) - \frac{n}{2}(\overline{X} - \mu)^\top \Sigma^{-1}(\overline{X} - \mu).$$

Maximizing this partially wrt μ gives

$$\ell(\Sigma) \propto -\frac{n}{2} \log \det(\Sigma) - \frac{n}{2} \operatorname{tr}(\widehat{\Sigma}\Sigma^{-1}).$$

Graphical lasso

Instead of solving $\arg\max_{\Sigma \succeq 0} \ell(\Sigma)$ directly (which would give $\widehat{\Sigma}$) we reformulate this problem in terms of the precision matrix K, that is

$$\widehat{K} = \arg\max_{K \succeq 0} \log \det(K) - \operatorname{tr}(\widehat{\Sigma}K)$$

Structure learning for general Gaussian graphical models

Recall the Gaussian log-density for data $X_1, \ldots X_n \sim N_p(\mu, \Sigma)$:

$$\ell(\mu, \Sigma) \propto -\frac{n}{2} \log \det(\Sigma) - \frac{n}{2} \operatorname{tr}(\widehat{\Sigma}\Sigma^{-1}) - \frac{n}{2}(\overline{X} - \mu)^\top \Sigma^{-1}(\overline{X} - \mu).$$

Maximizing this partially wrt μ gives

$$\ell(\Sigma) \propto -\frac{n}{2} \log \det(\Sigma) - \frac{n}{2} \operatorname{tr}(\widehat{\Sigma}\Sigma^{-1}).$$

Graphical lasso

Instead of solving $\arg\max_{\Sigma \succeq 0} \ell(\Sigma)$ directly (which would give $\widehat{\Sigma}$) we reformulate this problem in terms of the precision matrix K, that is

$$\widehat{K}_\lambda = \underset{K \succeq 0}{\arg\max} \log \det(K) - \operatorname{tr}(\widehat{\Sigma}K) - \lambda |K|_1$$

where $|K|_1 = \sum_{ij} |K_{ij}|$ and $\lambda \geq 0$ is a tuning parameter. This is a convex optimization problem that can be efficiently solved.

Graphical lasso

$$\widehat{K}_\lambda = \underset{K \succeq 0}{\arg\max} \log\det(K) - \mathrm{tr}(\widehat{\Sigma}K) - \lambda |K|_1$$

- The ℓ_1-penalization term $\lambda |K|_1$ penalizes the **complexity** of the concentration matrix.

Graphical lasso

$$\widehat{K}_\lambda = \arg\max_{K \succeq 0} \log\det(K) - \mathrm{tr}(\widehat{\Sigma}K) - \lambda|K|_1$$

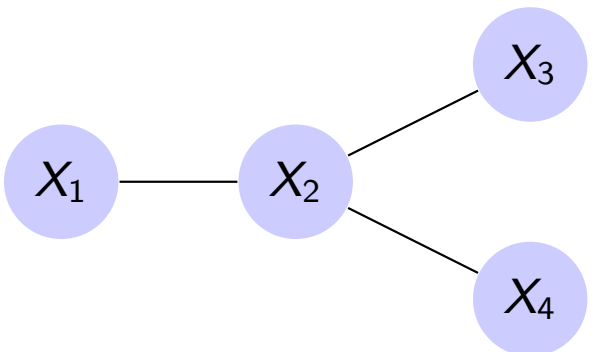
- The ℓ_1-penalization term $\lambda|K|_1$ penalizes the **complexity** of the concentration matrix.
- For increasing λ, one obtains simpler models and for decreasing λ models closer to the MLE $\widehat{K} = \widehat{\Sigma}^{-1}$.

Graphical lasso

$$\widehat{K}_\lambda = \underset{K \succeq 0}{\arg\max} \log\det(K) - \mathrm{tr}(\widehat{\Sigma}K) - \lambda |K|_1$$

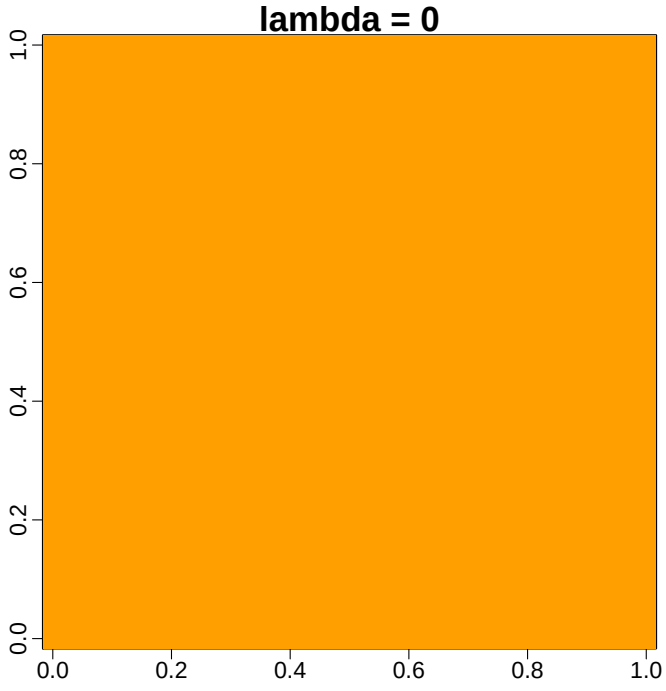
- The ℓ_1-penalization term $\lambda|K|_1$ penalizes the **complexity** of the concentration matrix.
- For increasing λ, one obtains simpler models and for decreasing λ models closer to the MLE $\widehat{K} = \widehat{\Sigma}^{-1}$.
- The geometry of the ℓ_1-penalization forces some of the entries of K to be **exactly zero**, and therefore one obtains a graphical model!

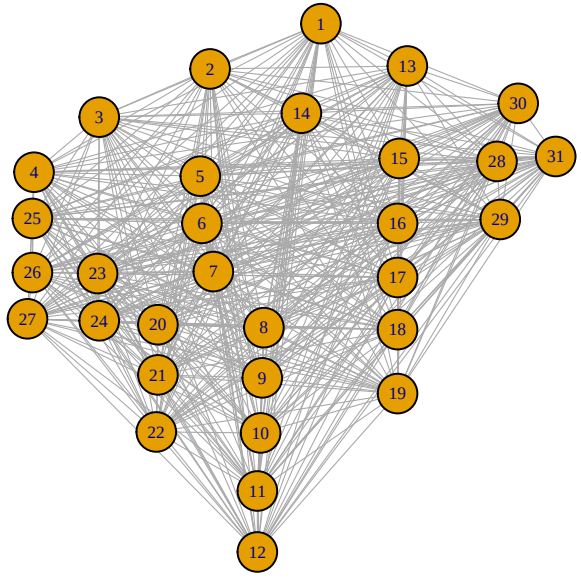
Example:



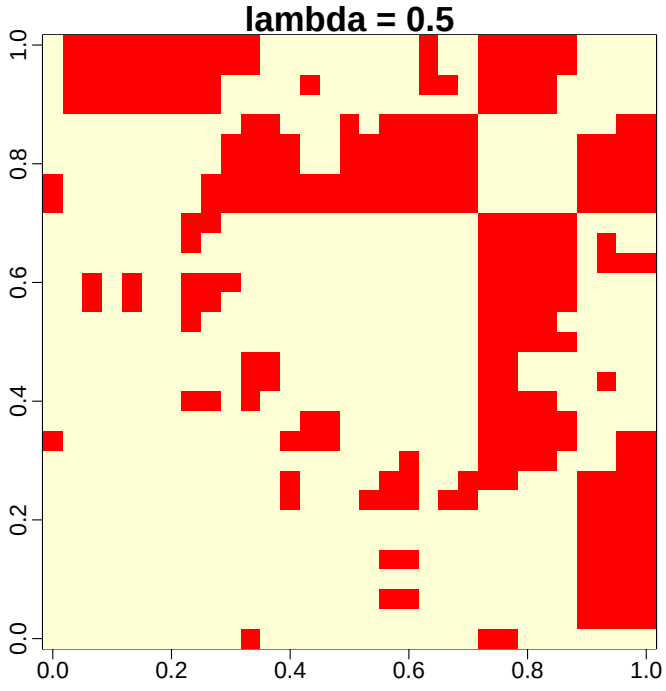
$$\widehat{K}_\lambda = \begin{bmatrix} * & * & 0 & 0 \\ * & * & * & * \\ 0 & * & * & 0 \\ 0 & * & 0 & * \end{bmatrix}$$

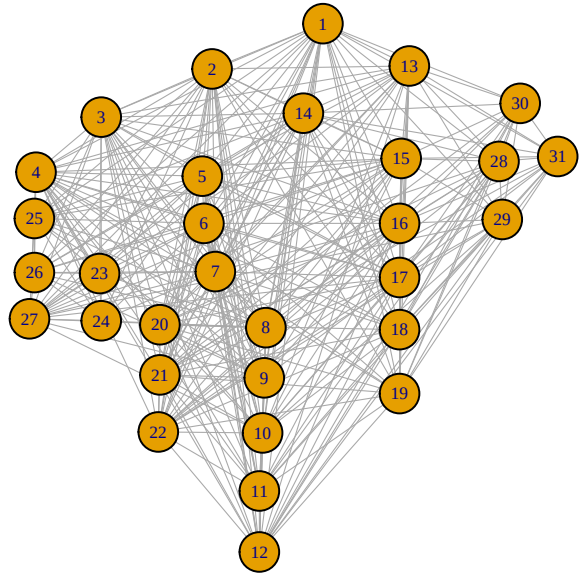
River discharge data: graphical lasso



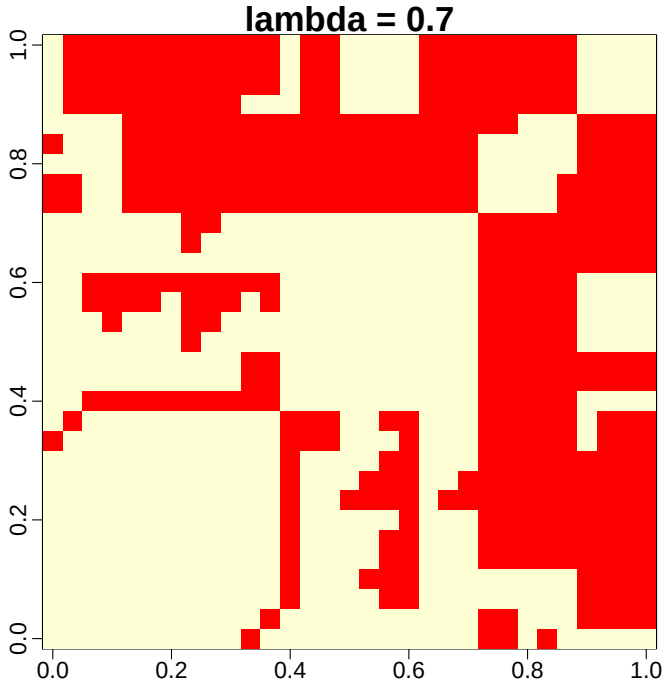


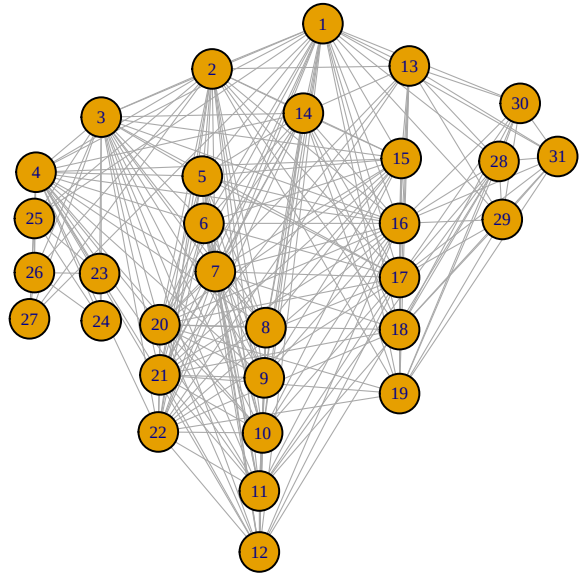
River discharge data: graphical lasso



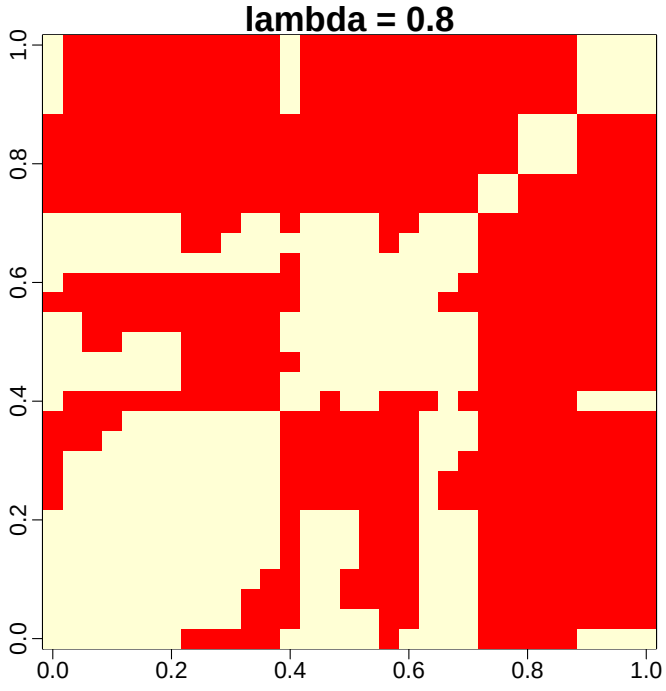


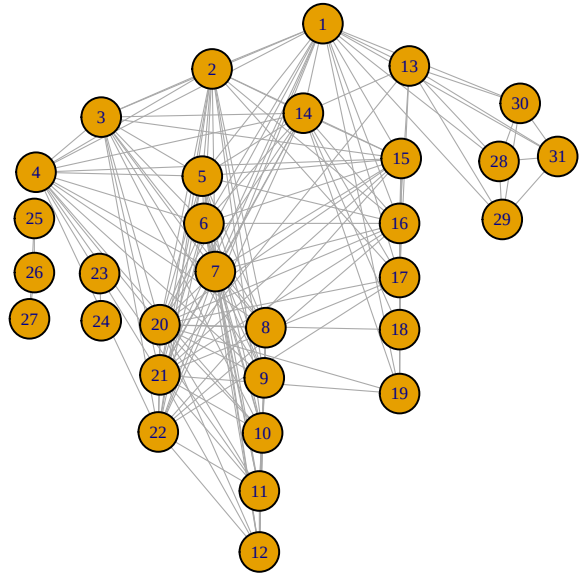
River discharge data: graphical lasso



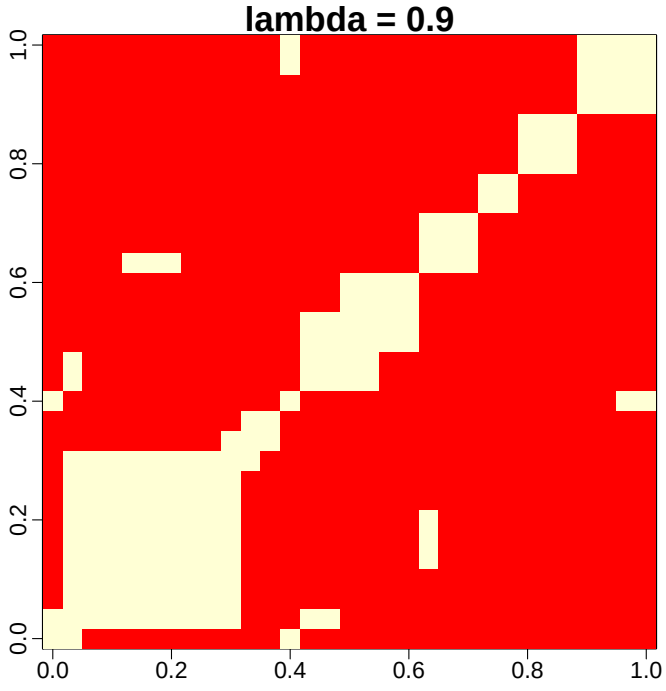


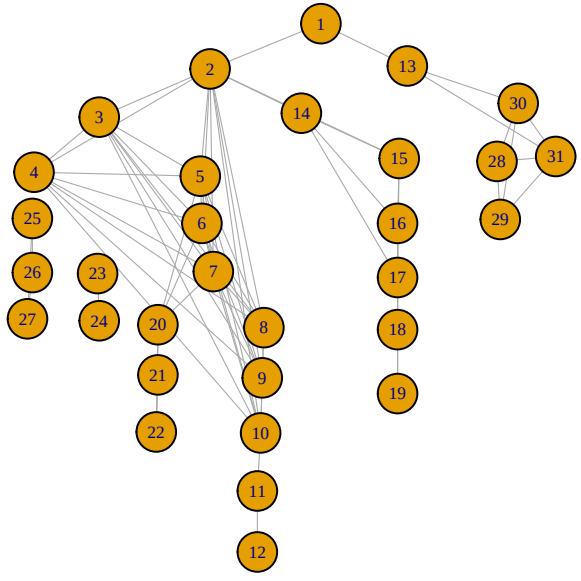
River discharge data: graphical lasso





River discharge data: graphical lasso





River discharge data: graphical lasso