

Multivariate Random Variables

Max Turgeon

STAT 7200–Multivariate Statistics

Joint distributions

- Let X and Y be two random variables.
- The *joint distribution function* of X and Y is

$$F(x, y) = P(X \leq x, Y \leq y).$$

- More generally, let Y_1, \dots, Y_p be p random variables. Their *joint distribution function* is

$$F(y_1, \dots, y_p) = P(Y_1 \leq y_1, \dots, Y_p \leq y_p).$$

Joint densities

- If F is absolutely continuous almost everywhere, there exists a function f called the *density* such that

$$F(y_1, \dots, y_p) = \int_{-\infty}^{y_1} \cdots \int_{-\infty}^{y_p} f(u_1, \dots, u_p) du_1 \cdots du_p.$$

- The *joint moments* are defined as follows:

$$E(Y_1^{n_1} \cdots Y_p^{n_p}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} u_1^{n_1} \cdots u_p^{n_p} f(u_1, \dots, u_p) du_1 \cdots du_p.$$

- **Exercise:** Show that this is consistent with the univariate definition of $E(Y_1^{n_1})$, i.e. $n_2 = \cdots = n_p = 0$.

Marginal distributions i

- From the joint distribution function, we can recover the *marginal distributions*:

$$F_i(x) = \lim_{\substack{y_j \rightarrow \infty \\ j \neq i}} F(y_1, \dots, y_p).$$

- More generally, we can find the joint distribution of a subset of variables by sending the other ones to infinity:

$$F(y_1, \dots, y_r) = \lim_{\substack{y_j \rightarrow \infty \\ j > r}} F(y_1, \dots, y_p), \quad r < p.$$

Marginal distributions ii

- Similarly, from the joint density function, we can recover the *marginal densities*:

$$f_i(x) = \int_{-\infty}^{\infty} f(u_1, \dots, u_p) du_1 \cdots \widehat{du_i} \cdots du_p.$$

- In other words, we are integrating *out* the other variables.

Example i

- Let $R = [a_1, b_1] \times \cdots \times [a_p, b_p] \subseteq \mathbb{R}^p$ be a hyper-rectangle, with $a_i < b_i$, for all i .
- If $\mathbf{Y} = (Y_1, \dots, Y_p)$ is **uniformly distributed** on R , then its density is given by

$$f(y_1, \dots, y_p) = \begin{cases} \prod_{i=1}^p \frac{1}{b_i - a_i} & (y_1, \dots, y_p) \in R, \\ 0 & \text{else.} \end{cases}$$

- For convenience, we can also use the indicator function:

$$f(y_1, \dots, y_p) = \prod_{i=1}^p \frac{I_{[a_i, b_i]}(y_i)}{b_i - a_i}.$$

Example ii

- We then have

$$\begin{aligned} F(y_1, \dots, y_p) &= \int_{-\infty}^{y_1} \cdots \int_{-\infty}^{y_p} f(u_1, \dots, u_p) du_1 \cdots du_p \\ &= \prod_{i=1}^p \left(\frac{y_i - a_i}{b_i - a_i} I_{[a_i, b_i]}(y_i) + I_{[b_i, \infty)}(y_i) \right). \end{aligned}$$

- Finally, note that we recover the *univariate* uniform distribution by sending all components but one to infinity:

$$F_i(x) = \lim_{\substack{y_j \rightarrow \infty \\ j \neq i}} F(y_1, \dots, y_p) = \frac{x - a_i}{b_i - a_i} I_{[a_i, b_i]}(x) + I_{[b_i, \infty)}(x).$$

Introduction to Copulas i

- **Copula theory** provides a general and powerful way to model general multivariate distributions.
- The main idea is that we can decouple (and recouple) the *marginal* distributions and the *dependency structure* between each component.
 - Copulas capture this dependency structure.
 - Sklar's theorem tells us about how to combine the two.

Definition

A p -dimensional copula is a function $C : [0, 1]^p \rightarrow [0, 1]$ that arises as the distribution function (CDF) of a random vector whose marginal distributions are all uniform on the interval $[0, 1]$.

In particular, we have

$$C(1, \dots, u_i, \dots, 1) = u_i, \quad u_i \in [0, 1].$$

Introduction to Copulas iii

Probability integral transform

If Y is a continuous (univariate) random variable with CDF

F_Y , then

$$F_Y(Y) \sim U(0, 1).$$

Proof

$$\begin{aligned} P(F_Y(Y) \leq x) &= P(Y \leq F_Y^{-1}(x)) \\ &= F_Y(F_Y^{-1}(x)) \\ &= x. \end{aligned}$$



Sklar's Theorem i

- Using the Probability integral transform, we can prove one part of Sklar's theorem.
- More precisely, let $\mathbf{Y} = (Y_1, \dots, Y_p)$ be a continuous random vector with CDF F , and let F_1, \dots, F_p be the CDFs of the marginal distributions.
- We know that $F_1(Y_1), \dots, F_p(Y_p)$ are uniformly distributed on $[0, 1]$, and therefore the CDF of their joint distribution is a copula C .

Sklar's Theorem ii

$$\begin{aligned}C(u_1, \dots, u_p) &= P(F_1(Y_1) \leq u_1, \dots, F_p(Y_p) \leq u_p) \\&= P(Y_1 \leq F_1^{-1}(u_1), \dots, Y_p \leq F_p^{-1}(u_p)) \\&= F(F_1^{-1}(u_1), \dots, F_p^{-1}(u_p)).\end{aligned}$$

- By taking $u_i = F_i(y_i)$, we get

$$F(y_1, \dots, y_p) = C(F_1(y_1), \dots, F_p(y_p)).$$

Sklar's Theorem iii

Theorem

Let $\mathbf{Y} = (Y_1, \dots, Y_p)$ be any random vector with CDF F , and let F_1, \dots, F_p be the CDFs of the marginal distributions. There exist a copula C such that

$$F(y_1, \dots, y_p) = C(F_1(y_1), \dots, F_p(y_p)). \quad (1)$$

If the marginal distributions are absolutely continuous, then C is unique.

Conversely, given a copula C and univariate CDFs F_1, \dots, F_p , then Equation 1 defines a valid CDF for a p -dimensional random vector.

Examples i

- **Gaussian copulas:** Let Φ be the CDF of the standard univariate normal distribution, and let Φ_{Σ} be the CDF of multivariate normal distribution with mean 0 and covariance matrix Σ . The Gaussian copula C_G is defined as

$$C_G(u_1, \dots, u_p) = \Phi_{\Sigma}(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_p)).$$

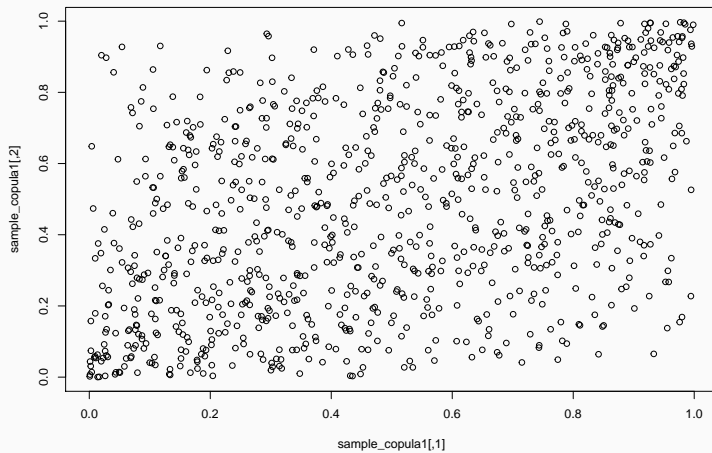
Examples ii

```
library(copula)

# Gaussian copula where correlation is 0.5
gaus_copula <- normalCopula(0.5, dim = 2)
sample_copula1 <- rCopula(1000, gaus_copula)

plot(sample_copula1)
```

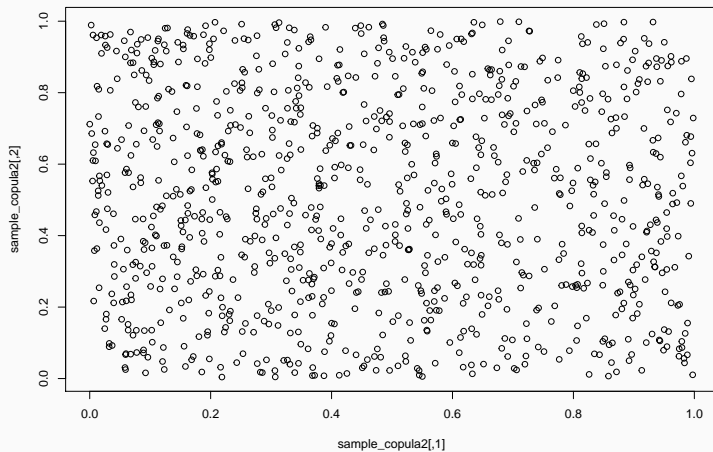
Examples iii



Examples iv

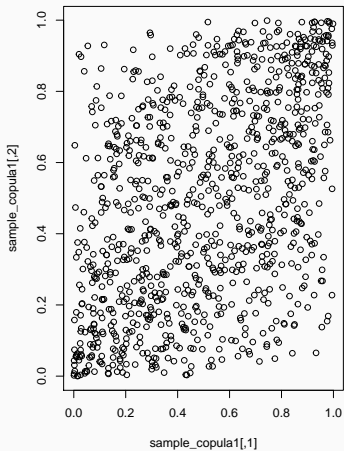
```
# Compare with independent copula,  
# i.e. two independent uniform variables.  
gaus_copula <- normalCopula(0, dim = 2)  
sample_copula2 <- rCopula(1000, gaus_copula)  
plot(sample_copula2)
```

Examples v

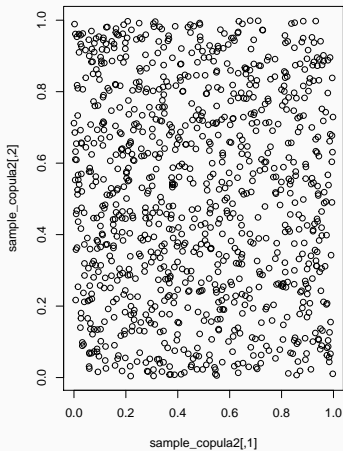


Examples vi

Corr. 0.5



Independent



Examples vii

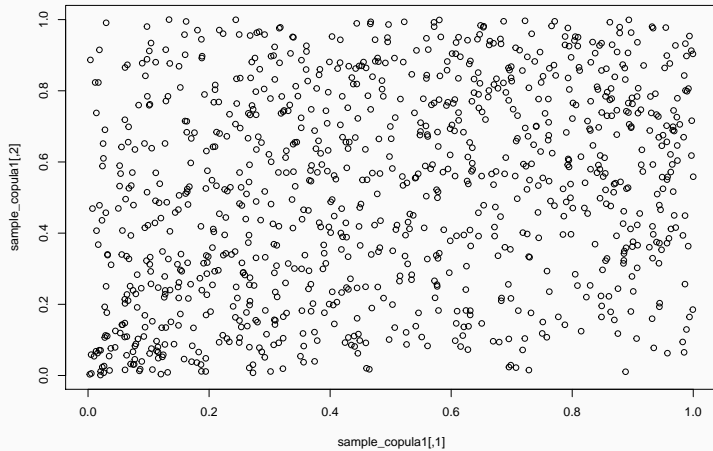
For a properly chosen θ :

Name	$C(u, v)$
Ali-Mikhail-Haq	$\frac{uv}{1-\theta(1-u)(1-v)}$
Clayton	$\max\left((u^{-\theta} + v^{-\theta} - 1)^{1/\theta}, 0\right)$
Independence	uv

Examples viii

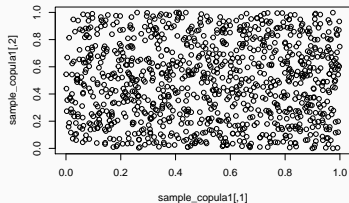
```
# Clayton copula with theta = 0.5  
clay_copula <- claytonCopula(param = 0.5)  
sample_copula1 <- rCopula(1000, clay_copula)  
  
plot(sample_copula1)
```

Examples ix

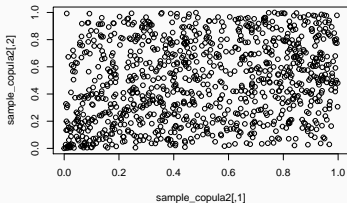


Examples x

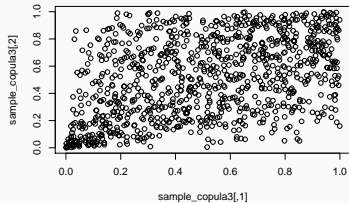
Independent



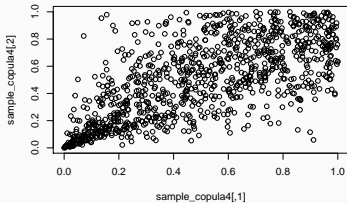
$\theta = 0.5$



$\theta = 1$



$\theta = 2$



Conditional distributions

- Let f_1, f_2 be the densities of random variables Y_1, Y_2 , respectively. Let f be the joint density.
- The *conditional density* of Y_1 given Y_2 is defined as

$$f(y_1|y_2) := \frac{f(y_1, y_2)}{f_2(y_2)},$$

whenever $f_2(y_2) \neq 0$ (otherwise it is equal to zero).

- Similarly, we can define the conditional density in $p > 2$ variables, and we can also define a conditional density for Y_1, \dots, Y_r given Y_{r+1}, \dots, Y_p .

Expectations

- Let $\mathbf{Y} = (Y_1, \dots, Y_p)$ be a random vector.
- Its *expectation* is defined entry-wise:

$$E(\mathbf{Y}) = (E(Y_1), \dots, E(Y_p)).$$

- **Observation:** The dependence structure has no impact on the expectation.

Covariance and Correlation i

- The multivariate generalization of the variance is the *covariance matrix*. It is defined as

$$\text{Cov}(\mathbf{Y}) = E \left((\mathbf{Y} - \mu)(\mathbf{Y} - \mu)^T \right),$$

where $\mu = E(\mathbf{Y})$.

- **Exercise:** The (i, j) -th entry of $\text{Cov}(\mathbf{Y})$ is equal to

$$\text{Cov}(Y_i, Y_j).$$

Covariance and Correlation ii

- Recall that we obtain the correlation from the covariance by dividing by the square root of the variances.
- Let V be the diagonal matrix whose i -th entry is $\text{Var}(Y_i)$.
 - In other words, V and $\text{Cov}(\mathbf{Y})$ have the same diagonal.
- Then we define the *correlation matrix* as follows:

$$\text{Corr}(\mathbf{Y}) = V^{-1/2} \text{Cov}(\mathbf{Y}) V^{-1/2}.$$

- **Exercise:** The (i, j) -th entry of $\text{Corr}(\mathbf{Y})$ is equal to

$$\text{Corr}(Y_i, Y_j).$$

Example i

- Assume that

$$\text{Cov}(\mathbf{Y}) = \begin{pmatrix} 4 & 1 & 2 \\ 1 & 9 & -3 \\ 2 & -3 & 25 \end{pmatrix}.$$

- Then we know that

$$V = \begin{pmatrix} 4 & 0 & 0 \\ 0 & 9 & 0 \\ 0 & 0 & 25 \end{pmatrix}.$$

Example ii

- Therefore, we can write

$$V^{-1/2} = \begin{pmatrix} 0.5 & 0 & 0 \\ 0 & 0.33 & 0 \\ 0 & 0 & 0.2 \end{pmatrix}.$$

- We can now compute the correlation matrix:

Example iii

$$\begin{aligned}\text{Corr}(\mathbf{Y}) &= \begin{pmatrix} 0.5 & 0 & 0 \\ 0 & 0.33 & 0 \\ 0 & 0 & 0.2 \end{pmatrix} \begin{pmatrix} 4 & 1 & 2 \\ 1 & 9 & -3 \\ 2 & -3 & 25 \end{pmatrix} \begin{pmatrix} 0.5 & 0 & 0 \\ 0 & 0.33 & 0 \\ 0 & 0 & 0.2 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0.17 & 0.2 \\ 0.17 & 1 & -0.2 \\ 0.2 & -0.2 & 1 \end{pmatrix}.\end{aligned}$$

Measures of Overall Variability

- In the univariate case, the variance is a scalar measure of spread.
 - In the multivariate case, the *covariance* is a matrix.
 - No easy way to compare two distributions.
 - For this reason, we have other notions of overall variability:
1. **Generalized Variance:** This is defined as the determinant of the covariance matrix.

$$GV(\mathbf{Y}) = \det(\text{Cov}(\mathbf{Y})).$$

2. **Total Variance:** This is defined as the trace of the covariance matrix.

$$TV(\mathbf{Y}) = \text{tr}(\text{Cov}(\mathbf{Y})).$$

Examples i

```
A <- matrix(c(5, 4, 4, 5), ncol = 2)

results <- eigen(A, symmetric = TRUE,
                 only.values = TRUE)

c("GV" = prod(results$values),
  "TV" = sum(results$values))

## GV TV
##  9 10
```


Examples ii

Compare this with the following

```
B <- matrix(c(5, -4, -4, 5), ncol = 2)
```

$GV(A) = 9$; $TV(A) = 10$

```
c("GV" = det(B),  
  "TV" = sum(diag(B)))
```

```
## GV TV
```

```
## 9 10
```

Measures of Overall Variability (cont'd)

- As we can see, we do lose some information:
 - In matrix B , we saw that the two variables are negatively correlated, and yet we get the same values
- But GV captures *some* information on dependence that TV does not.
 - Compare the following covariance matrices:

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}.$$

- *Interpretation:* A small value of the sampled Generalized Variance indicates either small scatter in data points or multicollinearity.

Geometric Interlude i

- A random vector \mathbf{Y} with positive definite covariance matrix Σ can be used to define a distance function on \mathbb{R}^p :

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

- This is called the *Mahalanobis distance* induced by Σ .
- **Exercise:** This indeed satisfies the definition of a distance:
 1. $d(x, y) = d(y, x)$
 2. $d(x, y) \geq 0$ and $d(x, x) = 0 \Leftrightarrow x = 0$
 3. $d(x, z) \leq d(x, y) + d(y, z)$

Geometric Interlude ii

- Using this distance, we can construct *hyper-ellipsoids* in \mathbb{R}^p as the set of all points x such that

$$d(x, 0) = 1.$$

- Equivalently:

$$x^T \Sigma^{-1} x = 1.$$

- Since Σ^{-1} is symmetric, we can use the spectral decomposition to rewrite it as:

$$\Sigma^{-1} = \sum_{i=1}^p \lambda_i^{-1} v_i v_i^T,$$

where $\lambda_1, \dots, \lambda_p$ are the eigenvalues of Σ .

Geometric Interlude iii

- We thus get a new parametrization if the hyper-ellipsoid:

$$\sum_{i=1}^p \left(\frac{v_i^T x}{\sqrt{\lambda_i}} \right)^2 = 1.$$

- **Theorem:** The volume of this hyper-ellipsoid is equal to

$$\frac{2\pi^{p/2}}{p\Gamma(p/2)} \sqrt{\lambda_1 \cdots \lambda_p}.$$

- In other words, the Generalized Variance is proportional to the square of the volume of the hyper-ellipsoid defined by the covariance matrix.
 - *Note:* the square root of the determinant of a matrix (if it exists) is sometimes called the *Pfaffian*.

Statistical Independence

- The variables Y_1, \dots, Y_p are said to be *mutually independent* if

$$F(y_1, \dots, y_p) = F(y_1) \cdots F(y_p).$$

- If Y_1, \dots, Y_p admit a joint density f (with marginal densities f_1, \dots, f_p), and equivalent condition is

$$f(y_1, \dots, y_p) = f(y_1) \cdots f(y_p).$$

- **Important property:** If Y_1, \dots, Y_p are mutually independent, then their joint moments factor:

$$E(Y_1^{n_1} \cdots Y_p^{n_p}) = E(Y_1^{n_1}) \cdots E(Y_p^{n_p}).$$

Linear Combination of Random Variables

- Let $\mathbf{Y} = (Y_1, \dots, Y_p)$ be a random vector. Let \mathbf{A} be a $q \times p$ matrix, and let $b \in \mathbb{R}^q$.
- Then the random vector $\mathbf{X} := \mathbf{A}\mathbf{Y} + b$ has the following properties:
 - **Expectation:** $E(\mathbf{X}) = \mathbf{A}E(\mathbf{Y}) + b$;
 - **Covariance:** $\text{Cov}(\mathbf{X}) = \mathbf{A}\text{Cov}(\mathbf{Y})\mathbf{A}^T$

Transformation of Random Variables i

- More generally, let $h : \mathbb{R}^p \rightarrow \mathbb{R}^p$ be a one-to-one function with inverse $h^{-1} = (h_1^{-1}, \dots, h_p^{-1})$. Define $\mathbf{X} = h(\mathbf{Y})$.
- Let J be the *Jacobian matrix* of h^{-1} :

$$\begin{pmatrix} \frac{\partial h_1^{-1}}{\partial y_1} & \dots & \frac{\partial h_1^{-1}}{\partial y_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial h_p^{-1}}{\partial y_1} & \dots & \frac{\partial h_p^{-1}}{\partial y_p} \end{pmatrix}.$$

- Then the density of \mathbf{X} is given by

$$g(x_1, \dots, x_p) = f(h_1^{-1}(x_1), \dots, h_p^{-1}(x_p)) |\det(J)|.$$

Transformation of Random Variables ii

- A few comments:
 - *This result is very useful for computing the density of transformations of normal random variables.*
 - If h is a linear transformation $\mathbf{Y} \mapsto A\mathbf{Y}$, then $J = A^{-1}$ (Exercise!).
 - See practice problems for further examples (or go back to your notes from mathematical statistics).

Characteristic function

- We will make use of the **characteristic function** φ_Y of a p -dimensional random vector \mathbf{Y} .
- The function $\varphi_Y : \mathbb{R}^p \rightarrow \mathbb{C}$ is defined as the expected value

$$\varphi_Y(\mathbf{t}) = E(\exp(i\mathbf{t}^T \mathbf{Y})),$$

where $i^2 = -1$.

- **Note:** The characteristic function of a random variable *always exists*.

Example 1 i

- Take the density of a normal distribution:

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

- Using the definition, we get

Example I ii

$$\begin{aligned}\varphi(t) &= \int_0^\infty \exp(itx) \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx \\ &= \int_0^\infty \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x^2 - 2\mu x + \mu^2 - 2it\sigma^2 x)}{2\sigma^2}\right) dx \\ &= \int_0^\infty \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x^2 - 2(\mu + it\sigma^2)x + \mu^2)}{2\sigma^2}\right) dx.\end{aligned}$$

Example I iii

- Let's complete the square:

$$\begin{aligned}x^2 - 2(\mu + it\sigma^2)x + \mu^2 &= \left(x - (\mu + it\sigma^2)\right)^2 \\&\quad + \left(\mu^2 - (\mu + it\sigma^2)^2\right) \\&= \left(x - (\mu + it\sigma^2)\right)^2 \\&\quad + \left(\mu^2 - (\mu^2 + 2it\mu\sigma^2 - (t\sigma^2)^2)\right) \\&= \left(x - (\mu + it\sigma^2)\right)^2 \\&\quad + \left((t\sigma^2)^2 - 2it\mu\sigma^2\right).\end{aligned}$$

Example I iv

- We thus get

$$\begin{aligned}\varphi(t) &= e^{\frac{-(t\sigma^2)^2 - 2it\mu\sigma^2}{2\sigma^2}} \int_0^\infty \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x - (\mu + it\sigma^2))^2}{2\sigma^2}\right) dx \\ &= \exp\left(-\frac{t^2\sigma^2}{2} + it\mu\right).\end{aligned}$$

Example II

- Take the density of a gamma distribution:

$$f(x; \alpha, \beta) = \frac{\beta^\alpha x^{\alpha-1} \exp(-\beta x)}{\Gamma(\alpha)}.$$

- Using the definition, we get

$$\begin{aligned}\varphi(t) &= \int_0^\infty \exp(itx) \frac{\beta^\alpha x^{\alpha-1} \exp(-\beta x)}{\Gamma(\alpha)} dx \\&= \frac{(\beta - it)^\alpha}{(\beta - it)^\alpha} \int_0^\infty \frac{\beta^\alpha x^{\alpha-1} \exp(-(\beta - it)x)}{\Gamma(\alpha)} dx \\&= \frac{\beta^\alpha}{(\beta - it)^\alpha} \int_0^\infty \frac{(\beta - it)^\alpha x^{\alpha-1} \exp(-(\beta - it)x)}{\Gamma(\alpha)} dx \\&= \left(1 - \frac{it}{\beta}\right)^{-\alpha}.\end{aligned}$$

Properties of the characteristic function i

1. $\varphi_Y(\mathbf{0}) = 1$
2. $|\varphi_Y(\mathbf{t})| \leq 1$ for all \mathbf{t}
3. $\varphi_Y(-\mathbf{t}) = \overline{\varphi_Y(\mathbf{t})}$
4. $\varphi_Y(\mathbf{t})$ is uniformly continuous.
5. Two random vectors are equal in distribution if and only if their characteristic functions are equal.
6. The components of $\mathbf{Y} = (Y_1, \dots, Y_p)$ are mutually independent if and only if $\varphi_Y(\mathbf{t}) = \prod_{i=1}^p \varphi_{Y_i}(t_i)$.

Properties of the characteristic function ii

Theorem

Let \mathbf{Y}_n be a sequence of p -dimensional random vectors, and let φ_n be the characteristic function of \mathbf{Y}_n . Then \mathbf{Y}_n converges in distribution to \mathbf{Y} if and only if the sequence φ_n converges pointwise to a function φ that is continuous at the origin. When this is the case, the function φ is the characteristic function of the limiting distribution \mathbf{Y} .

Cramer-Wold Theorem

Two random vectors \mathbf{X} and \mathbf{Y} are equal in distribution if and only if the linear combinations $\mathbf{t}^T \mathbf{X}$ and $\mathbf{t}^T \mathbf{Y}$ are equal in distribution for all vectors $\mathbf{t} \in \mathbb{R}^p$.

Proof

Let $\varphi_{\mathbf{X}}, \varphi_{\mathbf{Y}}$ be the characteristic functions of \mathbf{X} and \mathbf{Y} , respectively. Let $s \in \mathbb{R}$. Using the definition, we can see that

$$\varphi_{\mathbf{t}^T \mathbf{X}}(s) = E(\exp(is(\mathbf{t}^T \mathbf{X}))) = E(\exp(i(st)^T \mathbf{X})) = \varphi_{\mathbf{X}}(st).$$

The result follows from the uniqueness of characteristic functions. □