STA 437/2005:
Methods for Multivariate Data

Week 5: Non-Gaussian Distributions

Piotr Zwiernik

University of Toronto

Table of contents

Elliptical distributions

Why Study Elliptical Distributions?

- Generalize the multivariate normal distribution.
- Model data with heavy tails or outliers.
- Maintain symmetry and linear correlation structures.
- Applications in finance, insurance, and environmental studies.

Spherical Distributions

Orthogonal Matrices: $O(m) = \{U \in \mathbb{R}^{m \times m} : U^\top U = I_m\}$.

Spherical distribution

A random vector $X \in \mathbb{R}^m$ has a *spherical distribution* if for any $U \in O(m)$:

$$X \stackrel{d}{=} UX.$$

Characteristic function satisfies: $\psi_X(\boldsymbol{t}) = \psi_{UX}(\boldsymbol{t}) = \psi_X(U^\top \boldsymbol{t})$ and so **equivalently** $\psi_X(\boldsymbol{t})$ depends only on $\|\boldsymbol{t}\|$. Thus, the same applies to the density:

$$f_X(\mathbf{x}) \;=\; h(\|\boldsymbol{x}\|) \qquad \text{for some } h \text{ (generator)}.$$

Examples of Spherical Distributions

Standard normal distribution $Z \sim N_m(0, I_m)$ is a simple example.

Spherical scale mixture of normals

If $Z \sim N_m(0, I_m)$ and a random variable $\tau > 0$ is independent of Z, then:

$$X = \frac{1}{\sqrt{\tau}} Z$$

has a spherical distribution.

Indeed: Let $U \in O(m)$, then

$$UX = \frac{1}{\sqrt{\tau}} UZ \stackrel{d}{=} \frac{1}{\sqrt{\tau}} Z = X.$$

Moment Structure of Spherical Distributions

Spherical symmetry implies:

$$\begin{aligned} \mathbb{E}[X] &= 0, \\ \mathrm{var}(X) &= cI_m, \quad \text{for some } c \geq 0. \end{aligned}$$

Indeed: $\mathrm{var}(X) = \mathrm{var}(UX) = U\mathrm{var}(X)U^\top$ for any $U \in O(m)$

For $X = \frac{1}{\sqrt{\tau}}Z$ with $Z \sim N(0, I_m)$, $\tau > 0$, $\tau \perp\!\!\!\perp Z$:

$$\mathrm{var}(X) = \mathbb{E}[\tau^{-1}]I_m.$$

Indeed: $\mathbb{E}[X] = \mathbb{E}[\frac{1}{\sqrt{\tau}}Z] = \mathbb{E}[\frac{1}{\sqrt{\tau}}]\mathbb{E}[Z] = \mathbf{0}_m$ and so

$$\mathrm{var}(X) \;=\; \mathbb{E}XX^\top - \mathbb{E}[X]\mathbb{E}[X]^\top \;=\; \mathbb{E}[\tfrac{1}{\tau}ZZ^\top] \;=\; \mathbb{E}[\tfrac{1}{\tau}]\mathbb{E}[ZZ^\top] \;=\; \mathbb{E}[\tfrac{1}{\tau}]I_m$$

Independence of $\|X\|$ and $\frac{X}{\|X\|}$

Key Property

If X is spherical, the norm $\|X\| = \sqrt{X^\top X}$ is independent of the direction $\frac{X}{\|X\|}$.

Proof Sketch: Let $U \in O(m)$. Then:

$$\frac{X}{\|X\|} \overset{d}{=} \frac{UX}{\|UX\|} = U\frac{X}{\|X\|}.$$

The vector $\frac{X}{\|X\|}$ is rotationally invariant $\Longrightarrow$ has uniform distribution on the unit sphere (independent of what $\|X\|$ is).

A formal proof uses polar coordinates, see the notes.

Elliptical Distribution $E(\mu, \Sigma)$

Recall that $Z \sim N_m(\mathbf{0}_m, I_m)$ then $X = \mu + \Sigma^{1/2} Z \sim N_m(\mu, \Sigma)$.

Elliptical distribution

A random vector $X \in \mathbb{R}^m$ has an elliptical distribution if:

$$X = \mu + \Sigma^{1/2} Z,$$

where Z is a spherical random vector.

The density of $X \sim E(\mu, \Sigma)$ is of the form

$$f_X(\mathbf{x}) \;=\; c_m \sqrt{\det \Sigma^{-1}}\, h((\mathbf{x} - \mu)^\top \Sigma^{-1} (\mathbf{x} - \mu)).$$

The generator g controls the shape of the distribution (and its tails in particular).

Again: Why Elliptical Distributions?

- Generalize the multivariate normal distribution.
- Model data with heavy tails or outliers.
- Maintain symmetry and linear correlation structures.
- Applications in finance, insurance, and environmental studies.

Scale Mixtures of Normals

Scale mixture of normals is a special class of elliptical distributions.

Stochastic representation:

$$X = \mu + \frac{1}{\sqrt{\tau}}\Sigma^{1/2}Z,$$

where $Z \sim N_m(0, I_m)$ and $\tau > 0$ is independent of Z.

Special Cases of Scale Mixture of Normals

- $\tau \equiv 1$: Multivariate normal.
- $\tau \sim \frac{1}{k}\chi^2_k$: Multivariate t-distribution with k degrees of freedom.
  - Smaller k means heavier tails. Gaussian is the limit $k \to \infty$.
- $\tau \sim \text{Exp}(1)$: Multivariate Laplace.

Covariance and Correlation in Elliptical Distributions

Σ is called the **scale matrix**. It is generally not equal to the covariance matrix.

$$\mathrm{Var}(X) = c\Sigma, \quad c > 0.$$

Correlation structure is still governed by Σ:

$$R_{ij} \;=\; \frac{c\Sigma_{ij}}{\sqrt{c\Sigma_{ii}c\Sigma_{jj}}} \;=\; \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii}\Sigma_{jj}}}.$$

Similarly, if $X \sim E(\mu, \Sigma)$ and $X = (X_A, X_B)$ then

$$\mathbb{E}(X_A | X_B = \mathbf{x}_B) \;=\; \mathbb{E}(X_A) - \Sigma_{A,B}\Sigma_{B,B}^{-1}(\mathbf{x}_B - \mu_B)$$

exactly as in the Gaussian case.

Copula models

Cumulative Distribution Function (CDF)

Let $X = (X_1, \ldots, X_m)$ be a random vector. Its **CDF** is:

$$F(x_1, \ldots, x_m) = \mathbb{P}(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_m \leq x_m).$$

Marginal CDF: $F_1(x_1) = \mathbb{P}(X_1 \leq x_1) = \lim_{x_2 \to \infty} \cdots \lim_{x_m \to \infty} F(x_1, x_2, \ldots, x_m)$.
(similar for any other margin)

If f is the corresponding density, then:

$$f(x_1, \ldots, x_m) = \frac{\partial^m}{\partial x_1 \cdots \partial x_m} F(x_1, \ldots, x_m)$$

$$F(x_1, \ldots, x_m) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_m} f(y_1, \ldots, y_m) \mathrm{d}y_1 \cdots \mathrm{d}y_m.$$

If $U \sim U[0,1]$ then $F(u) = u$ $u \in [0,1]$.

What is a Copula?

- A **copula** is a function that captures the **dependence structure** between random variables, separate from their marginal distributions.

Definition

A function $C : [0,1]^m \to [0,1]$ is a **copula** if it is a CDF with uniform marginals, that is, $C_1(u_1) = u_1, \ldots, C_m(u_m) = u_m$, where C_i are the marginal CDF's.

For example, the copula $C(\boldsymbol{u}) = u_1 \cdots u_m$ corresponds to a m independent $U[0,1]$.

Why use copulas?

- To model non-Gaussian dependencies.
- To analyze dependence independently of marginal behaviors.

Sklar's Theorem

Theorem (Sklar, 1959)

Let $X = (X_1, \ldots, X_m)$ be a random vector with joint CDF F and marginals $F_1, \ldots, F_m$. There exists a unique copula C such that:

$$F(x_1, \ldots, x_m) = C(F_1(x_1), \ldots, F_m(x_m)).$$

Conversely, given marginals $F_1, \ldots, F_m$ and a copula C, the joint CDF F is defined by the same formula.

- C captures **dependence structure**.
- $F_1, \ldots, F_m$ capture marginal behaviors.

Understanding Sklar's Theorem

- When $m = 1$, $C(u) = u$, the identity function on $[0, 1]$.
- If X is continuous with CDF F, then $F(X) \sim U(0, 1)$.

  Proof: If X is continuous, F is strictly increasing on the support. Hence

$$\mathbb{P}(F(X) \leq u) = \mathbb{P}(X \leq F^{-1}(u)) = F(F^{-1}(u)) = u.$$

- Let $X = (X_1, \ldots, X_m)$ with CDF F.
- Define $U_i = F_i(X_i)$, where F_i are the marginal CDFs.
- The transformed variables $U = (U_1, \ldots, U_m)$ have uniform marginals.

$$\mathbb{P}(U_1 \leq u_1, \ldots, U_m \leq u_m) = C(u_1, \ldots, u_m).$$

- Also note that

$$\mathbb{P}(U_1 \leq u_1, \ldots, U_m \leq u_m) = F(F_1^{-1}(u_1), \ldots, F_m^{-1}(u_m)) \tag{1}$$

  and so the copula can be computed explicitly.
- Sklar's theorem ensures C is unique for continuous distributions.

Given copula C and margins F_i we easily get the corresponding distribution using (1):

$$F(x_1, \ldots, x_m) \;=\; C(F_1(x_1), \ldots, F_m(x_m)).$$

Simple Example of a Copula

- Joint CDF:

$$F_{X,Y}(x,y) = \begin{cases} 0 & x < 0 \text{ or } y < 0, \\ x^2 y^2 & 0 \le x, y \le 1, \\ 1 & x > 1 \text{ and } y > 1, \\ \min(x^2, y^2) & \text{otherwise.} \end{cases}$$

- Marginal CDFs:

$$F_X(x) = x^2, \quad F_Y(y) = y^2 \quad \text{for } 0 \le x, y \le 1.$$

- Copula:

$$C(u,v) = uv \quad \text{if } u, v \le 1.$$

Sampling

Fix a copula $C(\boldsymbol{u})$ and suppose we can sample from it.

Transform the copula sample

Consider a sample $\boldsymbol{u}^{(1)}, \ldots, \boldsymbol{u}^{(n)}$ from the copula.

Transform the data to have the right marginals $F_1, \ldots, F_m$:

$$\mathbf{x}_i^{(t)} := F_i^{-1}(\boldsymbol{u}_i^{(t)}) \qquad \text{for all } i = 1, \ldots, m, t = 1, \ldots, n.$$

The sample $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$ has the right marginals and the right dependence structure.

We will later show how to sample from some popular copula models.

Invariance under Monotone Transformations

Copulas are invariant under monotone transformations.

Consider $Y_i := f_i(X_i)$, where f_i are strictly increasing transformations. Then the copula of X is the same as the copula of Y.

Proof outline: Let G be the CDF of Y and G_i the marginal CDF of Y_i

- By (1), equiv. show $F(F_1^{-1}(u_1), \ldots, F_m^{-1}(u_m)) = G(G_1^{-1}(u_1), \ldots, G_m^{-1}(u_m))$
- $G_i(y_i) = \mathbb{P}(Y_i \leq y_i) = \mathbb{P}(f_i(X_i) \leq y_i) = \mathbb{P}(X_i \leq f_i^{-1}(y_i)) = F_i(f_i^{-1}(y_i))$ and, in particular, $G_i^{-1} = f_i \circ F_i^{-1}$.
- Thus, $\{Y_i \leq G_i^{-1}(u_i)\} = \{f(X_i) \leq f_i(F_i^{-1}(u_i))\} = \{X_i \leq F_i^{-1}(u_i)\}$ and so

$$G(G_1^{-1}(u_1), \ldots, G_m^{-1}(u_m)) = \mathbb{P}\left(\bigcap_{i=1}^{m} \{Y_i \leq G_i^{-1}(u_i)\}\right)$$

$$= \mathbb{P}\left(\bigcap_{i=1}^{m} \{X_i \leq F_i^{-1}(u_i)\}\right) = F(F_1^{-1}(u_1), \ldots, F_m^{-1}(u_m)).$$

Density of a Copula

The PDF of a copula C is obtained by differentiating its CDF:

$$c(\mathbf{u}) = \frac{\partial^m C(\mathbf{u})}{\partial u_1 \cdots \partial u_m}.$$

Recall $C(\boldsymbol{u}) = F(F_1^{-1}(u_1), \ldots, F_m^{-1}(u_m))$. Using the chain rule:

$$c(\mathbf{u}) = \frac{f(\mathbf{x})}{\prod_{i=1}^m f_i(x_i)}, \qquad \text{where } x_i = F_i^{-1}(u_i) \text{ for all } i$$

where f is the joint density and f_i are marginal densities.

e.g. $C(\boldsymbol{u}) = u_1 \cdots u_m$ is the CDF of independent $U_i \sim U(0,1)$. The density is uniform on $[0,1]^m$. Given margins f_i, we get $f(\mathbf{x}) = \prod_i f_i(x_i)$.

Gaussian Copula

Gaussian copula is derived from the multivariate normal distribution $X \sim N_m(\mu, \Sigma)$.

By monotone invariance, we can assume $\mathbb{E}X_i = 0$, $\mathrm{var}(X_i) = 1$

- $\mu = 0$, Σ is a correlation matrix,
- each $X_i \sim N(0,1)$.

Let Φ be the CDF of $N(0,1)$ with PDF ϕ. Let $f(\mathbf{x}; \Sigma)$ be the PDF of $N_m(\mathbf{0}, \Sigma)$.

The density of the Gaussian copula $C(\boldsymbol{u}; \Sigma)$

Using the general formula, we get:

$$c(\mathbf{u}; \Sigma) \;=\; \frac{f(\mathbf{x}; \Sigma)}{\prod_{i=1}^m \phi(x_i)} \;=\; \det(\Sigma)^{-1/2} \exp\left(-\frac{1}{2}\mathbf{x}^\top(\Sigma^{-1} - I_m)\mathbf{x}\right),$$

where $x_i = \Phi^{-1}(u_i)$.

Sampling from the Gaussian copula $C(\boldsymbol{u}; \Sigma)$

Sample $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(n)} \sim N_m(\mathbf{0}, \Sigma)$.

Transform $\boldsymbol{u}_i^{(t)} = \Phi(\mathbf{z}_i^{(t)})$ for all $i = 1, \ldots, m$ and $t = 1, \ldots, n$.

The sample $\boldsymbol{u}^{(1)}, \ldots, \boldsymbol{u}^{(n)}$ comes from the Gaussian copula.

As described earlier, we can now transform this sample to get arbitrary margins.

Steps to Estimate a Copula: normalize data

Given data $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$, start by fixing a copula model (e.g. Gaussian).

We assume the CDF F of the data satisfies $F(\mathbf{x}) = C(F_1(x_1), \ldots, F_m(x_m))$.

However, the margins F_i are not known!.

Given a sample $\mathbf{x}_i^{(1)}, \ldots, \mathbf{x}_i^{(n)}$ of X_i we compute the **empirical CDF**

$$\widehat{F}_i(x_i) \ := \ \frac{1}{n} \sum_{t=1}^{n} \mathbf{1}\{\mathbf{x}_i^{(t)} \leq x_i\}.$$

Transform, the data using the empirical CDFs

$$\boldsymbol{u}_i^{(t)} \ = \ \widehat{F}_i(\mathbf{x}_i^{(t)}).$$

This transforms the data matrix $\mathbf{X}$ to $\boldsymbol{U}$ with uniform marginals.

Steps to Estimate a Copula: Fit the copula family

In the next step, we fit the data to the given copula family.

Often this is done by maximizing the log-likelihood $\sum_{t=1}^{n} \log c(\boldsymbol{u}^{(t)})$.

In the case of the Gaussian copula $C(\boldsymbol{u}; \Sigma)$:

- Transform the data to standard Gaussian margins: $\mathbf{y}_i^{(t)} = \Phi^{-1}(\boldsymbol{u}_i^{(t)})$.
- Fit the Gaussian likelihood for $N_m(\mathbf{0}, \Sigma)$ with the sample covariance $S_n = \frac{1}{n}\mathbf{Y}^\top \mathbf{Y}$.

Steps to Estimate a Copula: Evaluate the fit

As the last step, compare the fitted copula model with the observed data. Check whether the copula captures the dependence structure accurately.

We can generate samples from the fitted Gaussian copula.

Applications of Copulas

- **Finance:** Modeling dependencies in asset returns.
- **Insurance:** Understanding risks in correlated claims.
- **Environmental Science:** Joint modeling of extreme events (e.g., floods).
- **Medical Statistics:** Modeling dependence in survival times.

Gaussian mixtures

Gaussian Mixture Models (GMMs)

Definition: A Gaussian Mixture Model assumes that the data is generated from a mixture of K Gaussian components:

$$f(\mathbf{x}) = \sum_{k=1}^{K} \pi_k N_m(\mathbf{x}; \mu_k, \Sigma_k),$$

where:

- π_k: mixture weights ($\pi_k \geq 0$, $\sum_{k=1}^{K} \pi_k = 1$),
- μ_k, Σ_k: mean vector and covariance matrix of component k.

Key Properties:

- Flexible model for multimodal data distributions.
- Each Gaussian component corresponds to a sub-population.
- Approximation improves as K increases (universal approximator for continuous distributions).

Why Use Gaussian Mixtures?

Gaussian Mixture Models (GMMs) are widely used because of their:

- **Flexibility:** Ability to model complex data distributions.
- **Multimodality:** Handles datasets with multiple clusters or modes.
- **Interpretability:** Each Gaussian component represents a sub-population with interpretable parameters.
- **Clustering Applications:** GMMs are a natural probabilistic method for clustering.

Special Case: For simplicity, in clustering, we often assume $\Sigma_k = \Sigma$ for all k.

Latent Variable Representation of GMMs

Latent Variable Model: GMMs can be written using a latent variable $Z \in \{1, \ldots, K\}$, where:

- $\mathbb{P}(Z = k) = \pi_k$ (mixing proportions),
- $\mathbf{X} \mid Z = k \sim N_m(\mu_k, \Sigma_k)$ (conditional distribution).

Marginal Distribution: The marginal distribution of $\mathbf{X}$ is:

$$f(\mathbf{x}) = \sum_{k=1}^{K} \pi_k N_m(\mathbf{x}; \mu_k, \Sigma_k).$$

Advantages of Latent Representation:

- Facilitates model estimation (e.g., Expectation-Maximization).
- Provides a natural interpretation for clustering and component membership.

Likelihood Inference for GMMs

Objective: Estimate the parameters $\theta = \{\pi_k, \mu_k, \Sigma_k\}_{k=1}^{K}$ by maximizing the likelihood:

$$\ell(\theta) = \sum_{i=1}^{n} \log f(\mathbf{x}_i) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \pi_k N_m(\mathbf{x}_i; \mu_k, \Sigma_k) \right).$$

Challenges:

- The likelihood function is multimodal and unbounded.
- Direct maximization is computationally difficult.

Solution: Use the **Expectation-Maximization (EM)** algorithm to iteratively estimate parameters using the latent variable representation.

EM Algorithm for GMMs

The EM algorithm alternates between two steps to estimate parameters:

1. **E-step:** Compute the posterior probabilities (responsibilities) for each component:

$$w_{ik} = \mathbb{P}(Z = k \mid \mathbf{X} = \mathbf{x}_i, \theta) = \frac{\pi_k N_m(\mathbf{x}_i; \mu_k, \Sigma_k)}{\sum_{l=1}^{K} \pi_l N_m(\mathbf{x}_i; \mu_l, \Sigma_l)}.$$

2. **M-step:** Update the parameters using the responsibilities:

$$\begin{aligned}
\pi_k &= \frac{1}{n} \sum_{i=1}^{n} w_{ik}, \\
\mu_k &= \frac{\sum_{i=1}^{n} w_{ik} \mathbf{x}_i}{\sum_{i=1}^{n} w_{ik}}, \\
\Sigma_k &= \frac{\sum_{i=1}^{n} w_{ik} (\mathbf{x}_i - \mu_k)(\mathbf{x}_i - \mu_k)^\top}{\sum_{i=1}^{n} w_{ik}}.
\end{aligned}$$

Convergence: Repeat until the log-likelihood $\ell(\theta)$ converges.

Example: GMM in Action

Simulated Data: Consider a 2D dataset generated from two Gaussian components:

$$\mu_1 = \begin{bmatrix} 2 \\ 2 \end{bmatrix}, \quad \Sigma_1 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}, \quad \pi_1 = 0.4,$$

$$\mu_2 = \begin{bmatrix} 7 \\ 7 \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} 1 & -0.3 \\ -0.3 & 1 \end{bmatrix}, \quad \pi_2 = 0.6.$$

- Use the EM algorithm to estimate π_k, μ_k, Σ_k.
- Visualize posterior probabilities after 1, 2, and 30 iterations.

Questions to Consider:

- How close are the estimated parameters to the true values?
- How does the choice of K affect results?

Visualization of EM Algorithm

- Initial parameter estimates result in poor cluster separation.
- Responsibilities evolve over iterations, improving separation.
- Log-likelihood increases monotonically with iterations.