STA 437/2005:
Methods for Multivariate Data

Week 5: Non-Gaussian Distributions

Piotr Zwiernik

University of Toronto

Modelling non-Gaussian distributions

Gaussian distribution has many properties that makes it very appealing.

It does however has some limitations:

- Problem with multimodal populations.
- Problem with asymetric distributions.
- Not suitable for modelling processes with extreme events.

Goal: Retain some of the advantages of the Gaussian removing some of its limitations.

We focus on three approaches:

- spherical and elliptical distributions
- copula modelling
- Gaussian mixtures

Table of contents

Elliptical distributions

Why Study Elliptical Distributions?

- Generalize the multivariate normal distribution.
- Model data with heavy tails or outliers.
  - higher probability of extreme events
- 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 $U \in O(m)$:

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

Example: $X \sim N_m(\mathbf{0}, I_m)$ or more generally $X \sim N_m(\mathbf{0}, \sigma^2 I_m)$.

Density generator and dependence on the norm

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 $\|\boldsymbol{t}\|$. The same applies to the density:

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

Examples of Spherical Distributions

The case $X \sim N_m(\mathbf{0}, \sigma^2 I_m)$ has a simple generalization.

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:

- $\mu = \mathbb{E}[X] = 0$,
- $\Sigma = \mathrm{var}(X) = cI_m$, for some $c \geq 0$.

Indeed: Let $\Sigma = U\Lambda U^\top$ be the spectral decomposition.

- $\Sigma = \mathrm{var}(X) = \mathrm{var}(VX) = V\mathrm{var}(X)V^\top = VU\Lambda U^\top V^\top$ for any $V \in O(m)$.
- take $V = U^\top$ to show that Σ must be diagonal, $\Sigma = \Lambda$.
- take V to be all the permutation matrices to conclude that $\Lambda = cI_m$.

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 $E(\mu, \Sigma)$ 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

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

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

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.

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 (particularly tractable subclass)

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.

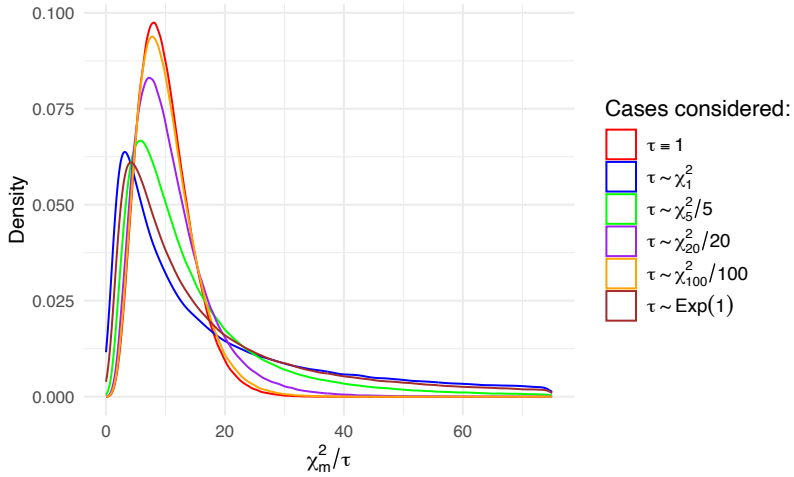
Special Cases of Scale Mixture of Normals

- $\tau \equiv 1$: 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.

Its about the tails (say $m = 10$)

For scale mixture of normals:

$$Y := \|X - \mu\|_\Sigma^2 \;=\; (X-\mu)^\top \Sigma^{-1}(X-\mu) \;=\; \tfrac{1}{\tau}\|Z\|^2 \;\stackrel{d}{=}\; \tfrac{1}{\tau}\chi_m^2.$$



Some tails are much heavier than Gaussian.

In the plot above we study $Y = \|X - \mu\|_{\Sigma}^2$ for X: normal, multivariate t, Laplace.

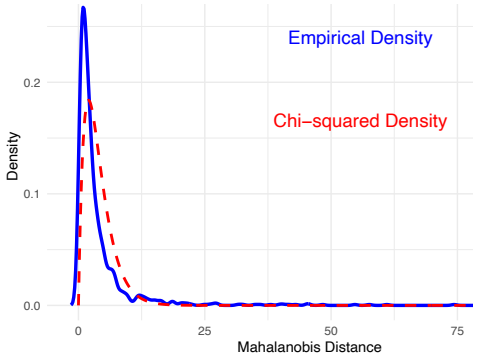
Case	$P(Y > 75)$	$P(Y > 500)$	$P(Y > 1000)$	$P(Y > 10000)$
Gaussian	0.000	0.000	0.000	0.000
t_{100}	0.000	0.000	0.000	0.000
t_{20}	0.000	0.000	0.000	0.000
t_5	0.019	0.000	0.000	0.000
Laplace	0.124	0.020	0.010	0.001
t_1	0.277	0.109	0.077	0.024

Table: Proportion of Samples Exceeding Thresholds

Simple illustration

In the notes we provide an example of four stocks: Apple, Microsoft, Google, Amazon.

Compare the empirical distribution of the Mahalanobis distance with χ^2_4 (Gaussian).



Empirical density seems to be more concentrated around zero.

But it has much heavier tails.

- $\mathbb{P}(\chi^2_4 > 20) \approx 0.$
- $\mathbb{P}(Y > 20) \approx 0.03$

This may be much more dramatic for smaller companies.

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 of X, 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 $F(u) = u$ for all $u \in [0,1]$.

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 **continuous** random vector with joint CDF F and marginals $F_1, \ldots, F_m$. There exists a unique copula C such that:

$$F(x_1, \ldots, x_m) = C(F_1(x_1), \ldots, F_m(x_m)). \tag{1}$$

Conversely, given marginals $F_1, \ldots, F_m$ and a copula C, F in (1) is a CDF of a multivariate distribution with given margins.

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

Understanding Sklar's Theorem

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 and margins F_i. Define $U_i := F_i(X_i)$.

- 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 $C(\boldsymbol{u})$ is given explicitly in terms of F and F_i's:

$$C(\boldsymbol{u}) = \mathbb{P}(F_1(X_1) \leq u_1, \ldots, F_m(X_m) \leq u_m) = F(F_1^{-1}(u_1), \ldots, F_m^{-1}(u_m)) \quad (2)$$

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.

$$\mathbb{P}(\mathbf{x}_i^{(t)} \leq s_i) = \mathbb{P}(F_i^{-1}(\boldsymbol{u}_i^{(t)}) \leq s_i) = \mathbb{P}(\boldsymbol{u}_i^{(t)} \leq F_i(s_i)) = F_i(s_i).$$

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

- By (2), equiv. show $F(F_1^{-1}(u_1), \ldots, F_m^{-1}(u_m)) = G(G_1^{-1}(u_1), \ldots, G_m^{-1}(u_m))$
- $G_i(y_i) = \mathbb{P}(Y_i \leq y_i) = \mathbb{P}(f_i(X_i) \leq y_i) = \mathbb{P}(X_i \leq f_i^{-1}(y_i)) = F_i(f_i^{-1}(y_i))$.
- Thus, $\{Y_i \leq G_i^{-1}(u_i)\} = \{F_i(f_i^{-1}(Y_i)) \leq u_i\} = \{F_i(X_i) \leq u_i\} = \{X_i \leq F_i^{-1}(u_i)\}$ 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 $C(\boldsymbol{u}) = F(F_1^{-1}(u_1), \ldots, F_m^{-1}(u_m))$. By chain rule and inverse function theorem:

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

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

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

Let Σ be a correlation matrix.

- 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.

The Gaussian copula model can still handle quite general distributions. Yet, it retains some of the computational advantages of the Gaussian distribution.

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** (proxy for F_i)

$$\widehat{F}_i(x_i) \;:=\; \frac{1}{n}\sum_{t=1}^{n} \mathbf{1}\{\mathbf{x}_i^{(t)} \le x_i\} \;\approx\; \mathbb{P}(X_i \le x_i) \;=\; F_i(x_i).$$

Transform, each row in the data matrix $\mathbf{X}$ 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

Mixture of Gaussians

We combine simple models into a complex model by taking a mixture of K multivariate Gaussian densities of the form:

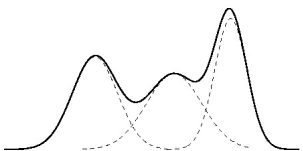
$$p(x) = \sum_{k=1}^{K} \pi_k N_m(x|\mu_k, \Sigma_k) \quad \text{for } x \in \mathbb{R}^m,$$

where $\pi_k \geq 0$, $\sum_{k=1}^{K} \pi_k = 1$, and $N_m(x|\mu_k, \Sigma_k)$ is the m-dim Gaussian density.

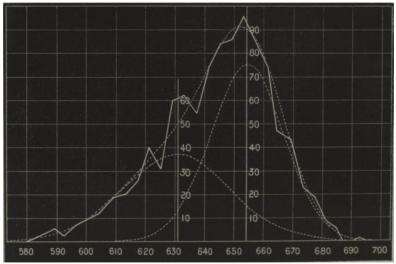
- Each Gaussian component has its own mean vector μ_k and covariance matrix Σ_k.
- The parameters π_k are called the mixing coefficients.

Example:

- $K = 3$ (three Gaussian components)
- $m = 1$ (univariate Gaussians)



The crabs from Naples bay



In 1892, scientists collected data on populations of the crab and observed that the ratio of forehead width to the body length actually showed a highly skewed distribution.

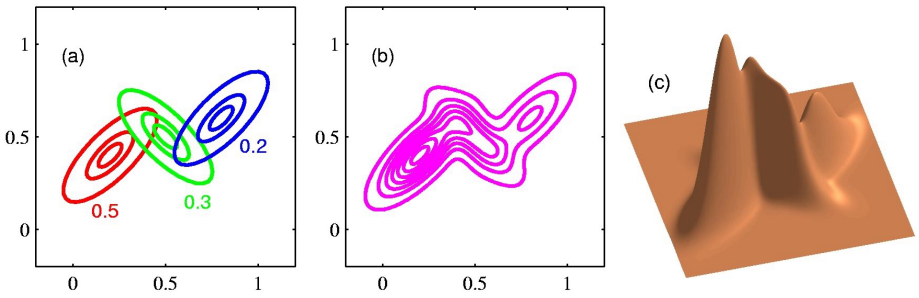
Source: *On Certain Correlated Variations in Carcinus maenas* (1893) W. F. Weldon.

They wondered whether this distribution could be the result of the population being a mix of two different normal distributions (two sub-species).

In **1894**, Karl Pearson proposed a method to fit this model (read here), whose modern version is the "method of moments". The method involved solving a higher order polynomial.

Mixture of Gaussians: 2D example

Illustration of a mixture of three Gaussians in 2D.



(a) Contours of constant density of each of the mixture components, along with the mixing coefficients.

(b) Contours of marginal density $p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k N_m(\mathbf{x}|\mu_k, \Sigma_k)$.

(c) A surface plot of the distribution $p(\mathbf{x})$.

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.

Mixture of Gaussians as a latent variable model

Recall: $p(x) = \sum_{k=1}^{K} \pi_k N_m(x|\mu_k, \Sigma_k)$.

- Consider a latent variable z with K states $z \in \{1, \ldots, K\}$.
- The distribution of z given by the mixing coefficients:

$$p(z = k) = \pi_k.$$

- Specify the conditional as $p(x|z = k) = N_m(x|\mu_k, \Sigma_k)$ with joint:

$$p(x, z = k) = p(z = k)p(x|z = k) = \pi_k N_m(x|\mu_k, \Sigma_k).$$

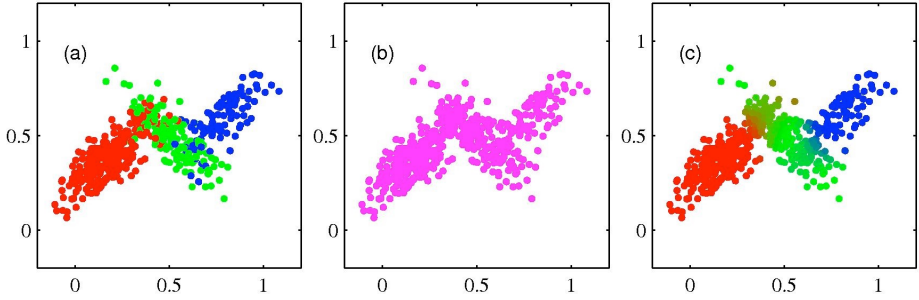
- Then the marginal $p(x)$ satisfies

$$p(x) = \sum_{k=1}^{K} p(x, z = k) = \sum_{k=1}^{K} \pi_k N_m(x|\mu_k, \Sigma_k).$$

Yet another illustration

The quantities $p(z|x)$ are called responsibilities.

Consider 500 points drawn from a mixture of three Gaussians.



Samples from the joint distribution p(x,z).

Samples from the marginal distribution p(x).

Same samples where colors represent the value of responsibilities.

The Likelihood function

Parameters: $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$, $\boldsymbol{\Sigma} = (\Sigma_1, \ldots, \Sigma_K)$.
Recall: $p(x|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{k=1}^{K} \pi_k N_m(x|\mu_k, \Sigma_k)$

- Represent the dataset $\{x_1, \ldots, x_N\}$ as $\boldsymbol{X} \in \mathbb{R}^{N \times m}$.
- The latent variable is represented by a vector $\boldsymbol{z} \in \mathbb{R}^N$.
- The log-likelihood takes the form

$$\log p(\boldsymbol{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k N_m(x_n|\mu_k, \Sigma_k) \right)$$

Maximum Likelihood (μ)

Recall: $\log p(\boldsymbol{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k N_m(x_n|\mu_k, \Sigma_k) \right)$.

- Differentiating wrt μ_k and setting to zero gives:

$$0 = \sum_{n=1}^{N} \frac{\pi_k N(x_n|\mu_k, \Sigma_k)}{\sum_j \pi_j N(x_n|\mu_j, \Sigma_j)} \Sigma_k^{-1}(x_n - \mu_k) = \sum_{n=1}^{N} p(z_n = k|x_n) \Sigma_k^{-1}(x_n - \mu_k)$$

$$= \Sigma_k^{-1} \left(\sum_{n=1}^{N} p(z_n = k|x_n) x_n - \mu_k \sum_{n=1}^{N} p(z_n = k|x_n) \right).$$

- Equivalently (as Σ_k is positive definite)

$$\mu_k = \sum_n \frac{p(z = k|x_n)}{N_k} x_n, \qquad N_k = \sum_n p(z = k|x_n).$$

- Simple interpretation: the MLE given by the weighted mean of the data weighted by the posterior $p(z = k|x_n)$.

Maximum Likelihood $(\boldsymbol{\Sigma}, \boldsymbol{\pi})$

Recall: $\log p(\boldsymbol{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k N_m(x_n|\mu_k, \Sigma_k) \right)$.

- Differentiating wrt Σ_k and setting to zero gives:

$$\Sigma_k = \sum_n \frac{p(z=k|x_n)}{N_k} (x_n - \mu_k)(x_n - \mu_k)^\top.$$

- Again data points weighted by posterior probabilities.

- Finally, for the weights π_k the MLE is

$$\pi_k = \frac{N_k}{\sum_{j=1}^{K} N_j} = \frac{N_k}{N}, \qquad N_k = \sum_n p(z=k|x_n).$$

Motivating the EM algorithm

- The MLE does not have a closed form solution.
- The estimates depend on the posterior probabilities $p(z = k|x_n)$, which themselves depend on those parameters.
- Indeed, recall that

$$p(z = k|x_n) = \frac{\pi_k N_m(x_n|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N_m(x_n|\mu_j, \Sigma_j)}.$$

- Iterative solution (EM algorithm):
  - Initialize the parameters to some values.

E-step Update the posteriors $p(z = k|x_n)$.

M-step Update model parameters $\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}$.

  - Repeat.

EM algorithm for Gaussian mixtures

- Initialize $\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}$.
- E-step: for each k, n compute the posterior probabilities

$$p(z = k|x_n) \;=\; \frac{\pi_k N_m(x_n|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N_m(x_n|\mu_j, \Sigma_j)}.$$

- M-step: Re-estimate model parameters

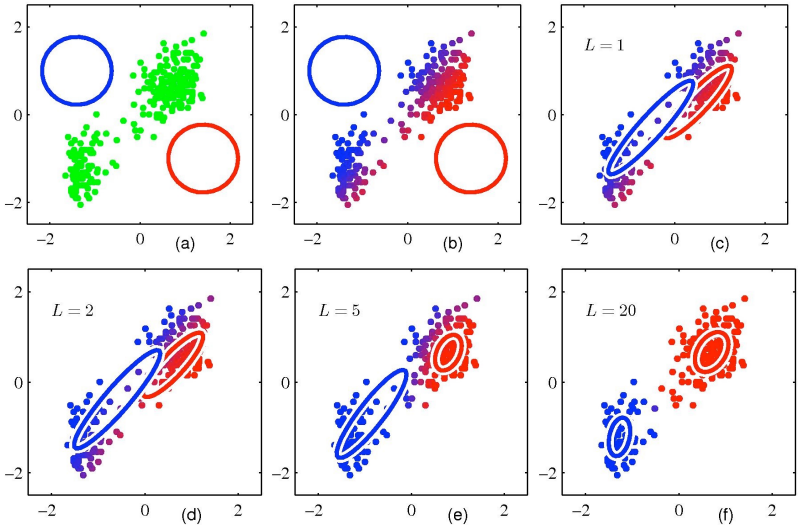
$$\mu_k^{\text{new}} = \sum_{n=1}^{N} \frac{p(z = k|x_n)}{N_k} x_n, \qquad N_k = \sum_{n=1}^{N} p(z = k|x_n),$$

$$\Sigma_k^{\text{new}} = \sum_{n=1}^{N} \frac{p(z = k|x_n)}{N_k} (x_n - \mu_k^{\text{new}})(x_n - \mu_k^{\text{new}})^\top,$$

$$\pi_k^{\text{new}} = \frac{N_k}{N}.$$

- Evaluate the log-likelihood and check for convergence.

Visualization of EM Algorithm



The General EM algorithm

Consider a general setting with latent variables.

- Observed dataset $\boldsymbol{X} \in \mathbb{R}^{N \times D}$, latent variables $\boldsymbol{Z} \in \mathbb{R}^{N \times K}$.

Maximize the log-likelihood $\log p(\boldsymbol{X}|\theta) = \log \left(\sum_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z}|\theta) \right)$.

- Initialize parameters θ^{old}.
- **E-step**: use θ^{old} to compute the posterior $p(\boldsymbol{Z}|\boldsymbol{X}, \theta^{\text{old}})$.
- **M-step**: $\theta^{\text{new}} = \arg\max_{\theta} Q(\theta, \theta^{\text{old}})$, where

$$Q(\theta, \theta^{\text{old}}) \;=\; \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X}, \theta^{\text{old}}) \log p(\boldsymbol{X}, \boldsymbol{Z}|\theta) \;=\; \mathbb{E}\Big(\log p(\boldsymbol{X}, \boldsymbol{Z}|\theta) \Big| \boldsymbol{X}, \theta^{\text{old}} \Big)$$

  which is tractable in many applications.

- Replace $\theta^{\text{old}} \leftarrow \theta^{\text{new}}$. Repeat until convergence.

Example: Gaussian mixture

- If z was observed, the MLE would be trivial

$$\log p(\boldsymbol{X}, \boldsymbol{Z}|\theta) = \sum_{n=1}^{N} \log p(x_n, z_n|\theta) = \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbb{1}(z_n = k) \log \left(\pi_k N(x_n|\mu_k, \Sigma_k)\right).$$

For the E-step: $p(\boldsymbol{Z}|\boldsymbol{X}, \theta) = \prod_{n=1}^{N} p(z_n|\boldsymbol{X}, \theta)$ we have

$$p(z_n = k|\boldsymbol{X}, \theta) = p(z_n = k|x_n, \theta) = \frac{\pi_k N_m(x_n|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N_m(x_n|\mu_j, \Sigma_j)}.$$

For the M-step: $\mathbb{E}(\mathbb{1}(z_n = k)|\boldsymbol{X}, \theta^{\text{old}}) = p(z_n = k|\boldsymbol{X}, \theta^{\text{old}})$ and so

$$\mathbb{E}\Big(\log p(\boldsymbol{X}, \boldsymbol{Z}|\theta) \Big| \boldsymbol{X}, \theta^{\text{old}} \Big) \;=\; \sum_{n=1}^{N} \sum_{k=1}^{K} p(z_n = k|\boldsymbol{X}, \theta^{\text{old}}) \log \left(\pi_k N(x_n|\mu_k, \Sigma_k)\right).$$

Maximizing gives the formulas on Slide 43.