Signal Processing

Lecture 15: Expectation Maximization & Mixture Models

Konstantinos Chatzilygeroudis - costashatz@upatras.gr

Department of Electrical and Computer Engineering
University of Patras

Template made by Panagiotis Papagiannopoulos

Motivation: Why Latent-Variable Models?

Many signal models involve **hidden / latent variables** that we do not observe directly.

Example (mixture / multi-regime signals):

$$y_i \text{ is generated by one of } K \text{ sources or regimes} \quad \Rightarrow \quad z_i \in \{1, \ldots, K\}.$$

Complete-data model:

$$p(y_i, z_i \mid \boldsymbol{\theta}) = p(z_i \mid \boldsymbol{\theta})\, p(y_i \mid z_i, \boldsymbol{\theta}).$$

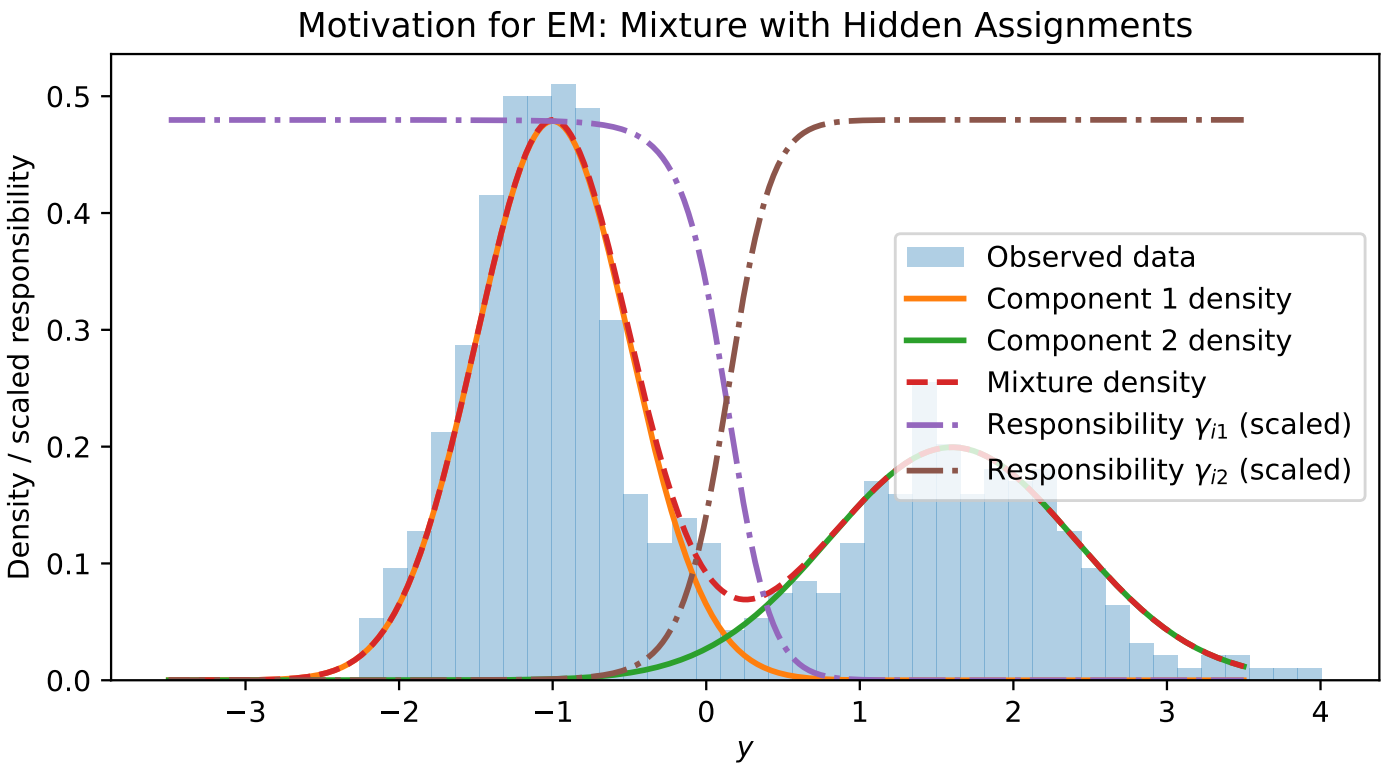
But we only observe y_i (N samples):

$$p(\mathbf{y} \mid \boldsymbol{\theta}) = \prod_{i=1}^{N} \sum_{k=1}^{K} p(z_i = k \mid \boldsymbol{\theta})\, p(y_i \mid z_i = k, \boldsymbol{\theta}).$$

Key difficulty:

$$\log p(\mathbf{y} \mid \boldsymbol{\theta}) = \sum_i \log \sum_k (\cdot) \quad \text{(log of sum is hard to optimize/compute).}$$

Motivation: Why Latent-Variable Models? (2)



Expectation Maximization idea: alternate between

- **E-step:** infer soft assignments $p(z_i = k \mid y_i, \boldsymbol{\theta})$,
- **M-step:** update $\boldsymbol{\theta}$ using these assignments.

Latent-Variable Models

We assume that the data is generated with the help of **unobserved (latent) variables**.

Observed data:

$$\boldsymbol{Y} = \{y_i\}_{i=1}^{N}$$

Latent variables:

$$\boldsymbol{Z} = \{z_i\}_{i=1}^{N} \quad \text{(e.g., cluster label, regime index, source ID).}$$

Generative structure:

$$p(\boldsymbol{Y}, \boldsymbol{Z} \mid \boldsymbol{\theta}) = p(\boldsymbol{Z} \mid \boldsymbol{\theta})\, p(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{\theta}).$$

Examples in signal processing

- Mixture / multi-regime signals: z_i = which regime generated y_i.
- Clustering features: z_i = cluster assignment.
- Speech / audio: z_i = phoneme/state producing a frame.

Complete-Data vs Incomplete-Data Likelihood

Complete-data likelihood (if we knew $\boldsymbol{Z}$):

$$p(\boldsymbol{Y}, \boldsymbol{Z} \mid \boldsymbol{\theta}) = \prod_{i=1}^{N} p(z_i \mid \boldsymbol{\theta})\, p(y_i \mid z_i, \boldsymbol{\theta}).$$

Incomplete-data likelihood (what we actually have):

$$p(\boldsymbol{Y} \mid \boldsymbol{\theta}) = \sum_{\boldsymbol{Z}} p(\boldsymbol{Y}, \boldsymbol{Z} \mid \boldsymbol{\theta}) \qquad (\text{or } \int d\boldsymbol{Z} \text{ if continuous}).$$

Key point:

Latent variables $\boldsymbol{Z}$ are missing $\Rightarrow$ we must marginalize them out.

Goal: estimate parameters by maximizing the incomplete-data likelihood

$$\boxed{\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{Y} \mid \boldsymbol{\theta}) \quad \Leftrightarrow \quad \arg\max_{\boldsymbol{\theta}} \log p(\boldsymbol{Y} \mid \boldsymbol{\theta}).}$$

Why the Likelihood is Hard

The log-likelihood for incomplete data becomes

$$\log p(\boldsymbol{Y} \mid \boldsymbol{\theta}) = \log \sum_{\boldsymbol{Z}} p(\boldsymbol{Y}, \boldsymbol{Z} \mid \boldsymbol{\theta}).$$

Problem: log of a sum

$$\log \sum_{\boldsymbol{Z}} (\cdot) \neq \sum_{\boldsymbol{Z}} \log(\cdot).$$

Consequences

- The objective is often **non-convex**.
- Direct maximization is typically **intractable**.
- Gradients involve ratios of sums of exponentials (unstable / messy).

We need a trick to handle the hidden $\boldsymbol{Z}$.

A Concrete Example: Mixture Likelihood

Suppose each y_i comes from one of K components:

$$p(z_i = k) = \pi_k, \qquad p(y_i \mid z_i = k, \boldsymbol{\theta}) = p_k(y_i \mid \boldsymbol{\theta}_k).$$

Marginal likelihood per sample:

$$p(y_i \mid \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \, p_k(y_i \mid \boldsymbol{\theta}_k).$$

Dataset log-likelihood:

$$\log p(\boldsymbol{Y} \mid \boldsymbol{\theta}) = \sum_{i=1}^{N} \log \Big(\sum_{k=1}^{K} \pi_k \, p_k(y_i \mid \boldsymbol{\theta}_k) \Big).$$

Expectation-Maximization (EM): Core Idea

We want to maximize the incomplete-data log-likelihood:

$$\log p(\boldsymbol{Y} \mid \boldsymbol{\theta}) = \log \sum_{\boldsymbol{Z}} p(\boldsymbol{Y}, \boldsymbol{Z} \mid \boldsymbol{\theta}).$$

Difficulty: the hidden $\boldsymbol{Z}$ is inside a log-sum.

EM key idea:

- Introduce a tractable lower bound on $\log p(\boldsymbol{Y} \mid \boldsymbol{\theta})$.
- Alternate between:
  - estimating latent variables (E-step),
  - updating parameters (M-step).

EM Derivation I: Introduce an Auxiliary Distribution

Let $q(\boldsymbol{Z})$ be *any* distribution over latent variables.
Rewrite the likelihood:

$$\log p(\boldsymbol{Y} \mid \boldsymbol{\theta}) = \log \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \, \frac{p(\boldsymbol{Y}, \boldsymbol{Z} \mid \boldsymbol{\theta})}{q(\boldsymbol{Z})}.$$

Interpretation:

$$\sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \, (\cdot) = \mathbb{E}_q[(\cdot)].$$

So

$$\log p(\boldsymbol{Y} \mid \boldsymbol{\theta}) = \log \mathbb{E}_q \left[\frac{p(\boldsymbol{Y}, \boldsymbol{Z} \mid \boldsymbol{\theta})}{q(\boldsymbol{Z})} \right].$$

EM Derivation I: Introduce an Auxiliary Distribution

Let $q(\boldsymbol{Z})$ be *any* distribution over latent variables.
Rewrite the likelihood:

$$\log p(\boldsymbol{Y} \mid \boldsymbol{\theta}) = \log \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \, \frac{p(\boldsymbol{Y}, \boldsymbol{Z} \mid \boldsymbol{\theta})}{q(\boldsymbol{Z})}.$$

Interpretation:

$$\sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \, (\cdot) = \mathbb{E}_q[(\cdot)].$$

So

$$\log p(\boldsymbol{Y} \mid \boldsymbol{\theta}) = \log \mathbb{E}_q \left[\frac{p(\boldsymbol{Y}, \boldsymbol{Z} \mid \boldsymbol{\theta})}{q(\boldsymbol{Z})} \right].$$

Jensen's Inequality:

- **Convex case (Jensen):** If φ is convex and X integrable, then
$$\varphi\big(\mathbb{E}[X]\big) \;\leqslant\; \mathbb{E}\big[\varphi(X)\big].$$
- **Concave case:** If φ is concave and X integrable, then the inequality reverses:
$$\varphi\big(\mathbb{E}[X]\big) \;\geqslant\; \mathbb{E}\big[\varphi(X)\big],$$
with equality iff X is a.s. constant or φ is affine on the support of X.

EM Derivation II: Jensen's Inequality (Lower Bound)

Apply Jensen's inequality (since $\log(\cdot)$ is concave):

$$\log \mathbb{E}_q[f(\boldsymbol{Z})] \;\geqslant\; \mathbb{E}_q[\log f(\boldsymbol{Z})].$$

Here, $f(\boldsymbol{Z}) = \frac{p(\boldsymbol{Y},\boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})}$. Therefore:

$$\boxed{\log p(\boldsymbol{Y} \mid \boldsymbol{\theta}) \geqslant \mathbb{E}_q\left[\log \frac{p(\boldsymbol{Y}, \boldsymbol{Z} \mid \boldsymbol{\theta})}{q(\boldsymbol{Z})}\right] = \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \log \frac{p(\boldsymbol{Y}, \boldsymbol{Z} \mid \boldsymbol{\theta})}{q(\boldsymbol{Z})}.}$$

Define the lower bound:

$$\boxed{\mathcal{L}(q, \boldsymbol{\theta}) = \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \log p(\boldsymbol{Y}, \boldsymbol{Z} \mid \boldsymbol{\theta}) - \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \log q(\boldsymbol{Z}).}$$

Two terms:

- Expected complete-data log-likelihood.
- Entropy of $q(\boldsymbol{Z})$.

EM Derivation III: Tightness via KL Divergence

We can prove that:

$$\log p(\boldsymbol{Y} \mid \boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{KL}(q(\boldsymbol{Z}) \parallel p(\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{\theta})) .$$

Since $\mathrm{KL}(\cdot\|\cdot) \geqslant 0$:

$$\mathcal{L}(q, \boldsymbol{\theta}) \leqslant \log p(\boldsymbol{Y} \mid \boldsymbol{\theta}).$$

Bound becomes tight when

$$q(\boldsymbol{Z}) = p(\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{\theta}).$$

KL divergence: measures how different q is from p (i.e., expected information loss using q instead of p).

$$\mathrm{KL}(q\|p) = \mathbb{E}_q\left[\log \frac{q(\boldsymbol{Z})}{p(\boldsymbol{Z})}\right] = \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \log \frac{q(\boldsymbol{Z})}{p(\boldsymbol{Z})} \;\geqslant\; 0,$$

with equality iff $q = p$ (a.e.).

Expectation-Maximization Algorithm

EM performs coordinate ascent on $\mathcal{L}(q, \boldsymbol{\theta})$:

E-step (update latent distribution):

$$\boxed{q^{(t+1)}(\boldsymbol{Z}) = p(\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{\theta}^{(t)}).}$$

Tightens the bound by minimizing KL: sets q to the posterior given our current parameters, $\boldsymbol{\theta}^{(t)}$, and data $\boldsymbol{Y}$.

M-step (update parameters):

$$\boxed{\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{q^{(t+1)}} \left[\log p(\boldsymbol{Y}, \boldsymbol{Z} \mid \boldsymbol{\theta})\right].}$$

Maximizes the bound w.r.t. parameters.

EM: Guarantees and Practical Notes

Monotonic improvement

$$\log p(\boldsymbol{Y} \mid \boldsymbol{\theta}^{(t+1)}) \geqslant \log p(\boldsymbol{Y} \mid \boldsymbol{\theta}^{(t)}).$$

What EM guarantees

- Each iteration *does not decrease* the incomplete likelihood.
- Converges to a **local optimum** (not necessarily global).

Practical notes

- Initialization matters (often use k-means / random restarts).
- Stop when likelihood improvement is small.
- Non-convexity $\Rightarrow$ multiple runs recommended.

Gaussian Mixture Models (GMM): Model

We model data $\{\boldsymbol{y}_i\}_{i=1}^N$, $\boldsymbol{y}_i \in \mathbb{R}^d$ as coming from a mixture of K Gaussians.

Latent assignment:

$$z_i \in \{1, \ldots, K\}.$$

Mixing weights:

$$p(z_i = k \mid \boldsymbol{\theta}) = \pi_k, \qquad \pi_k \geqslant 0, \ \sum_{k=1}^{K} \pi_k = 1.$$

Component densities:

$$p(\boldsymbol{y}_i \mid z_i = k, \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{y}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

Parameters:

$$\boldsymbol{\theta} = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K.$$

GMM: Incomplete (Observed) Likelihood

Marginal likelihood per sample:

$$p(\boldsymbol{y}_i \mid \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{y}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

Incomplete-data log-likelihood:

$$\log p(\boldsymbol{Y} \mid \boldsymbol{\theta}) = \sum_{i=1}^{N} \log \Big(\sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{y}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \Big).$$

Hard part: log-sum prevents direct optimization $\Rightarrow$ use EM.

GMM: Complete-Data Likelihood

If the assignments z_i were known:

$$p(\boldsymbol{Y}, \boldsymbol{Z} \mid \boldsymbol{\theta}) = \prod_{i=1}^{N} \prod_{k=1}^{K} \Big(\pi_k \, \mathcal{N}(\boldsymbol{y}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \Big)^{\mathbb{I}[z_i = k]}.$$

Complete-data log-likelihood:

$$\log p(\boldsymbol{Y}, \boldsymbol{Z} \mid \boldsymbol{\theta}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \mathbb{I}[z_i = k] \Big(\log \pi_k + \log \mathcal{N}(\boldsymbol{y}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \Big).$$

EM replaces $\mathbb{I}[z_i = k]$ by its posterior expectation, where $\mathbb{I}[z_i = k]$ is the **indicator function**:

$$\mathbb{I}[z_i = k] = \begin{cases} 1, & \text{if } z_i = k, \\ 0, & \text{otherwise.} \end{cases}$$

E-step: Responsibilities

E-step: compute posterior probability that sample i belongs to component k:

$$\gamma_{ik} \equiv p(z_i = k \mid \boldsymbol{y}_i, \boldsymbol{\theta}^{(t)}) = \frac{\pi_k^{(t)} \mathcal{N}(\boldsymbol{y}_i \mid \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)})}{\sum_{j=1}^{K} \pi_j^{(t)} \mathcal{N}(\boldsymbol{y}_i \mid \boldsymbol{\mu}_j^{(t)}, \boldsymbol{\Sigma}_j^{(t)})}.$$

Interpretation:

- $\gamma_{ik} \in [0, 1]$, and $\sum_k \gamma_{ik} = 1$.
- "Soft" cluster assignment (not hard labels).

M-step: Update Mixing Weights and Means

Let the effective number of points in cluster k be:

$$N_k = \sum_{i=1}^{N} \gamma_{ik}.$$

Mixing weights:

$$\boxed{\pi_k^{(t+1)} = \frac{N_k}{N}.}$$

Means:

$$\boxed{\boldsymbol{\mu}_k^{(t+1)} = \frac{1}{N_k} \sum_{i=1}^{N} \gamma_{ik} \boldsymbol{y}_i.}$$

Interpretation: re-weighted averages using responsibilities.

M-step: Update Covariances

Covariances:

$$\boxed{\boldsymbol{\Sigma}_k^{(t+1)} = \frac{1}{N_k} \sum_{i=1}^{N} \gamma_{ik} \, (\boldsymbol{y}_i - \boldsymbol{\mu}_k^{(t+1)})(\boldsymbol{y}_i - \boldsymbol{\mu}_k^{(t+1)})^{\mathsf{T}}.}$$

Interpretation: responsibility-weighted covariance within each component.

Numerical note: often regularize to avoid singularities:

$$\boldsymbol{\Sigma}_k \leftarrow \boldsymbol{\Sigma}_k + \epsilon \boldsymbol{I}.$$

GMM-EM Algorithm (Summary)

Initialize $\{\pi_k^{(0)}, \boldsymbol{\mu}_k^{(0)}, \boldsymbol{\Sigma}_k^{(0)}\}$.

Repeat until convergence:

- **E-step:** compute responsibilities γ_{ik}.
- **M-step:** update $\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$ using γ_{ik}.

Convergence check:

$$\log p(\boldsymbol{Y} \mid \boldsymbol{\theta}^{(t+1)}) - \log p(\boldsymbol{Y} \mid \boldsymbol{\theta}^{(t)}) < \delta.$$

Each iteration increases (or leaves unchanged) the incomplete likelihood.

Practical Notes for GMMs

Initialization matters

- Initial centers, $\boldsymbol{\mu}_k^{(0)}$, are very important!.
- Multiple random restarts help avoid bad local maxima.

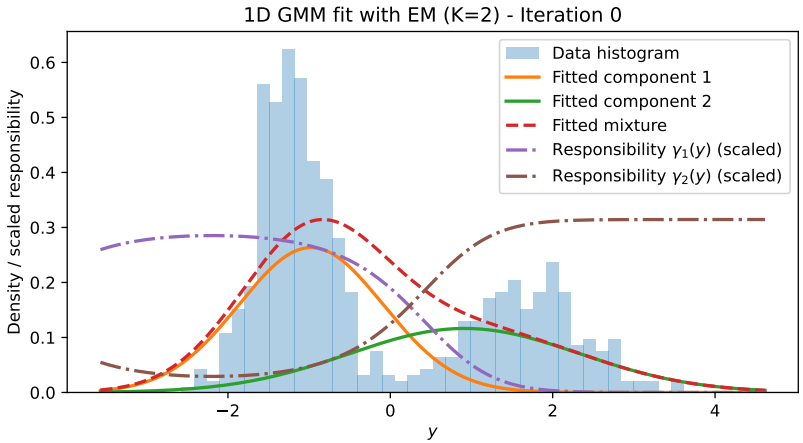
Choosing K

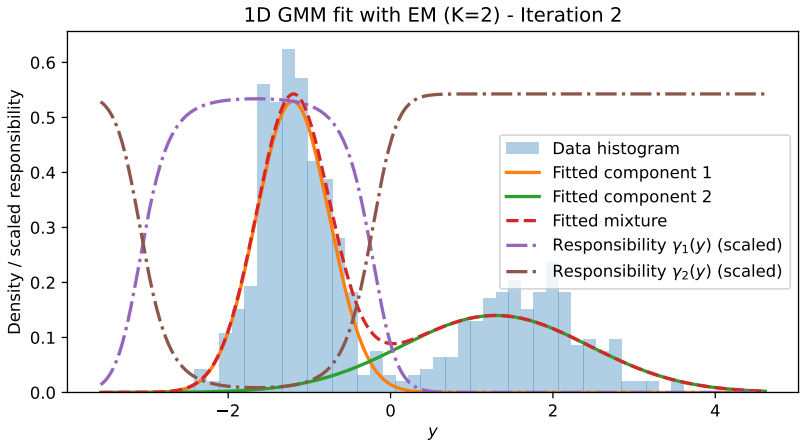
- BIC / AIC / cross-validation.
- Too small K: underfit; too large K: overfit.

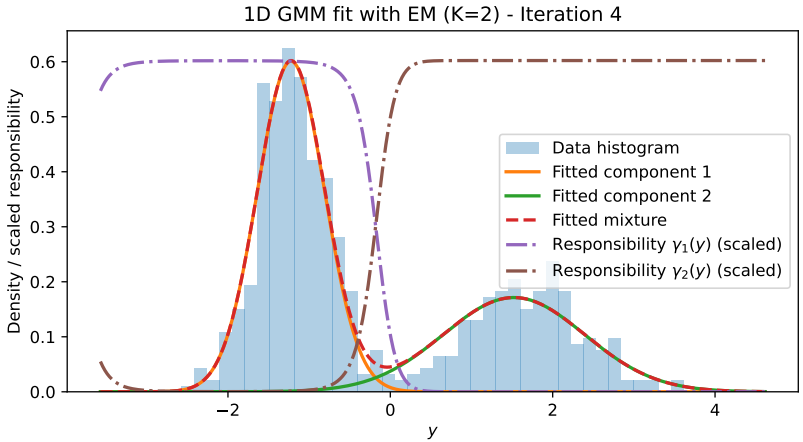
Covariance structure

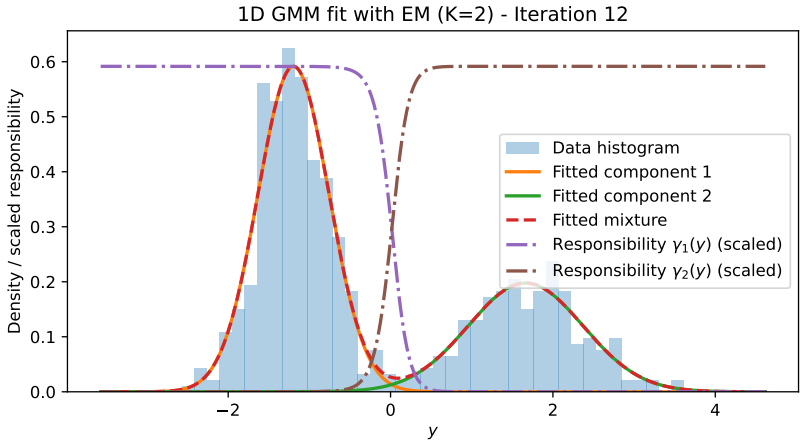
- Full $\boldsymbol{\Sigma}_k$: flexible but more parameters.
- Diagonal / spherical: faster, more stable.

GMMs: 1D Example

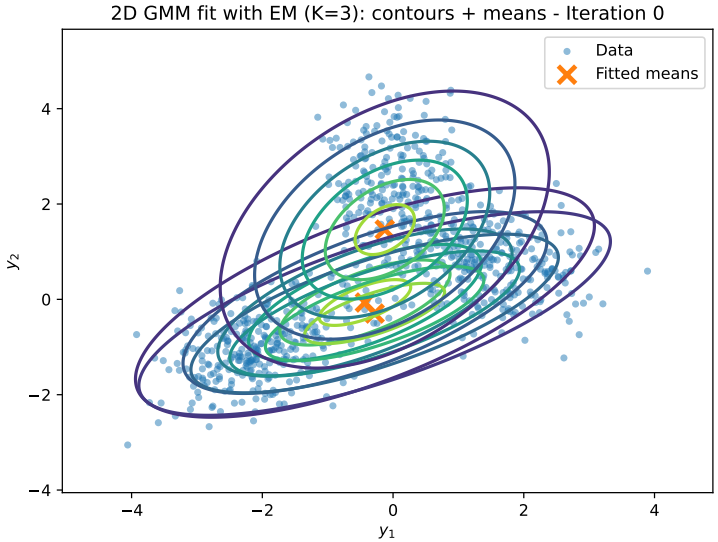


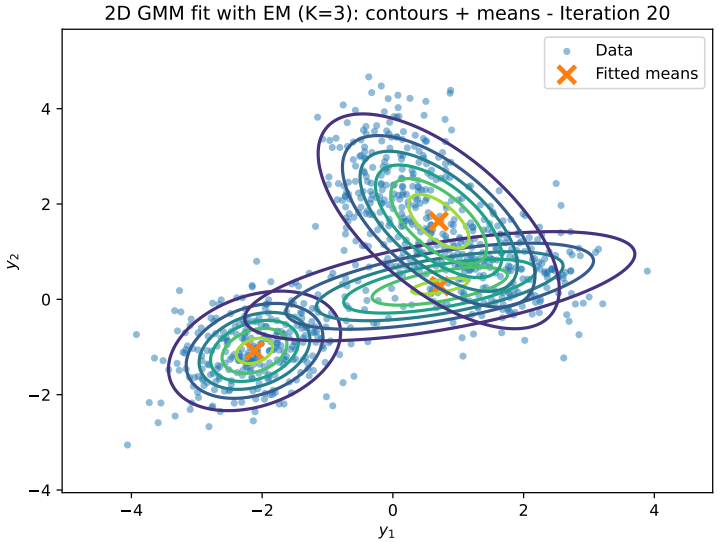


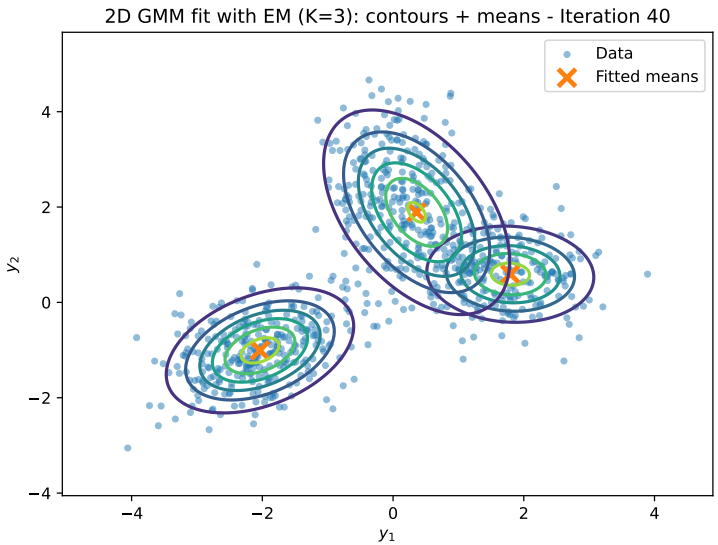


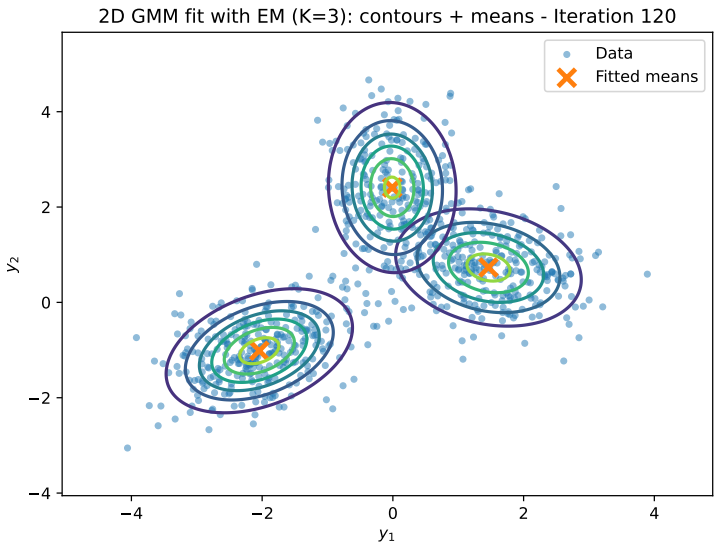


GMMs: 2D Example

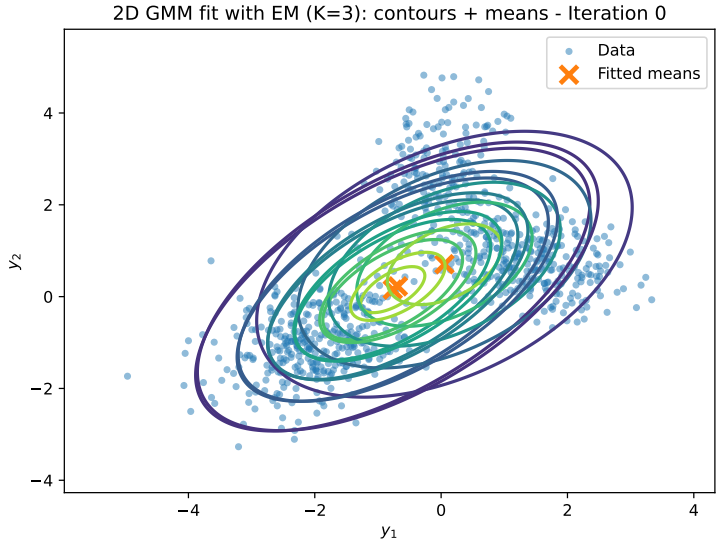


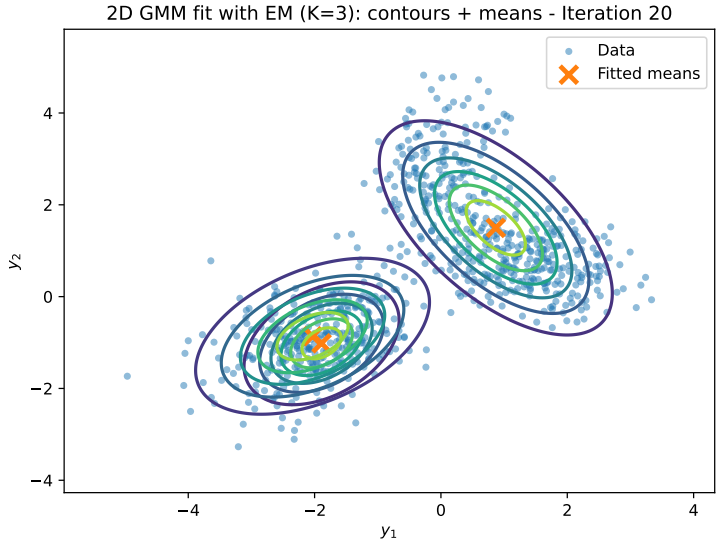


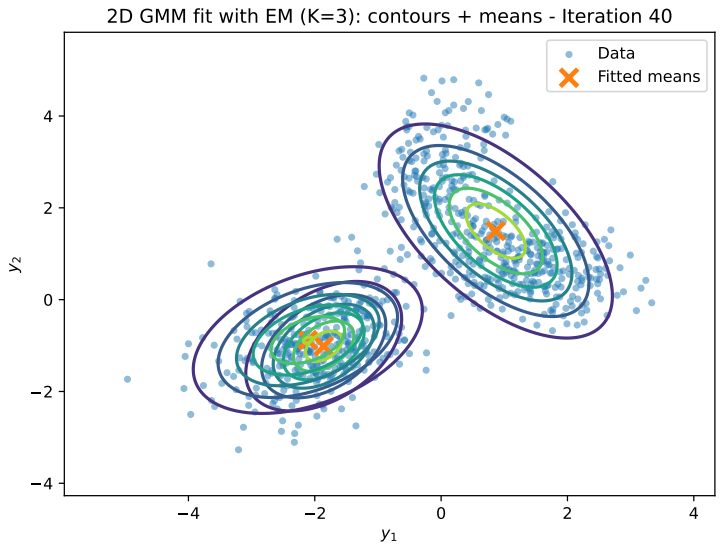


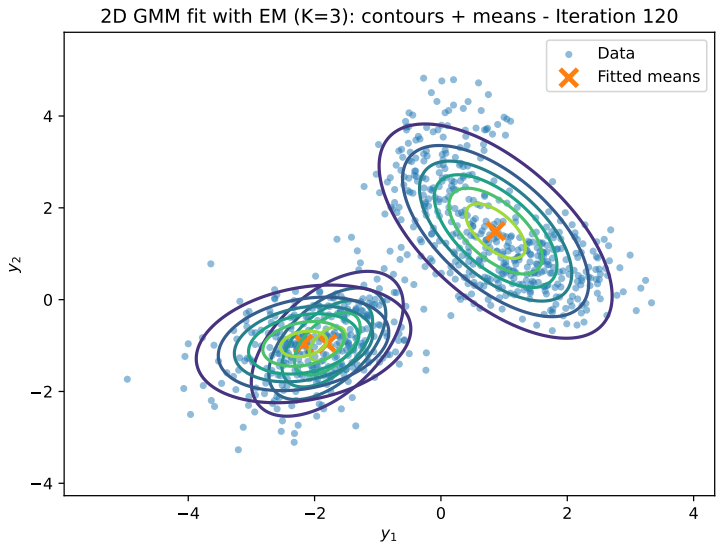


GMMs: 2D Example - Bad Initialization









Thank you

- **Any Questions?**
- **Office Hours**:
  - **Tue & Thu** *(09:00-11:00)*
  - 24/7 by email (`costashatz@upatras.gr`, subject: *ECE_SP_AM*)
- **Material and Announcements**

Laboratory of Automation & Robotics