Yale University

Latent Variables and EM

S&DS 265 · Introductory Machine Learning · Lecture 16

Zhuoran Yang

Fall 2026

AI usage: figures were generated, and these slides polished, with the help of AI tools.

Outline

1. Latent-Variable Mixture Models
2. The Posterior Assignment
3. Expectation–Maximization Updates
4. The Evidence Lower Bound
5. Checking the Fit, and What Can Go Wrong

Learning Objectives

In this lecture you will learn:

1. Why a latent component variable makes a mixture easy to describe but hard to fit;
2. How Bayes' rule turns current parameters into a posterior over the hidden assignment;
3. How weighted complete-data fitting produces the EM parameter updates;
4. Why the evidence lower bound explains EM's monotone likelihood improvement;
5. How restarts, effective counts, and covariance checks reveal mixture failures.

Outline

Motivating Example: What Kinds of Car Are These?

You are handed 392 cars, each with a weight and a mileage figure.

What kinds of car are in there, and how common is each kind?

Nothing in the table says which kind a car is.

A **kind** is a group of similar cars; in the model it will be called a component.

Example and data: ISLP `Auto` data (James et al., 2023).

The Auto Data

model	mpg	cylinders	horsepower	weight	year
chevrolet chevelle malibu	18	8	130	3504	70
buick skylark 320	15	8	165	3693	70
plymouth satellite	18	8	150	3436	70
amc rebel sst	16	8	150	3433	70

The local Auto table contains 392 complete fuel-economy records. Throughout the lecture the model sees only weight and mpg; every other column is held back.

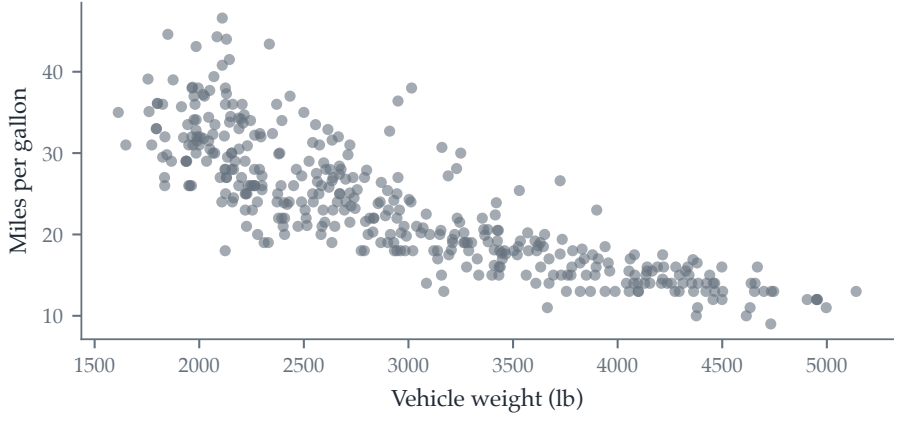
Source: First four complete rows of Auto.csv from the ISLP resources.

Reading One Row

model	mpg	cylinders	horsepower	weight	year
Chevelle Malibu	18	8	130	3504	70

This car achieves 18 mpg, has eight cylinders and 130 horsepower, weighs 3,504 pounds, and is from model year 1970. Which class of car it belongs to is a question the row does not answer.

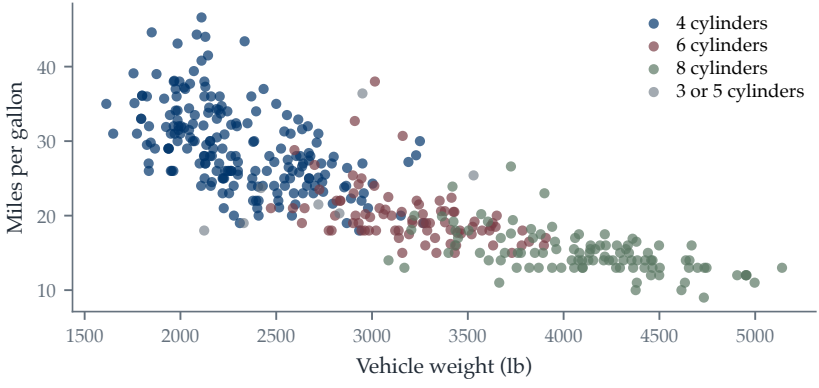
This Is Everything the Method Is Given



Each point is one car, $x_i \in \mathbb{R}^2$. Heavy cars use more fuel, so the cloud slopes down — but nothing in the picture says group, and no column in the table does either.

ISLP Auto data, 392 complete cases; weight and fuel economy only.

This Data Has an Answer Key



Auto happens to record the number of cylinders, and it lines up with the picture: three engine classes in different regions, overlapping where they meet. We will not let the model see this column — it is how we check the answer afterwards. In most problems no such column exists.

ISLP Auto data; cylinders are used to colour this figure only, never to fit a model.

A Hard Label Overstates What We Know

Last lecture's answer was one label per car, $c_i \in \{1, \dots, K\}$.

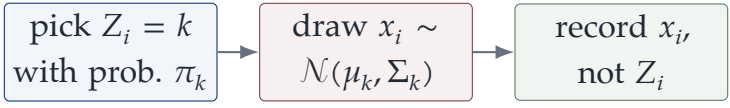
The Ford Maverick weighs 3,012 lb and gets 24 mpg, which puts it squarely between two of the groups. K-means files it under one of them — with the same confidence it gives a 1,800 lb compact.

We want the answer to be a share instead: this car is about half of one kind and half of the other. A share has to come from somewhere, so each group will need a spread and a prevalence, not just a center.

The Model: a Gaussian Mixture

Assume each car comes from one of $K = 3$ populations. Population k is Gaussian, and it is chosen with probability π_k:

$$\Pr(Z_i = k) = \pi_k, \qquad x_i \mid Z_i = k \;\sim\; \mathcal{N}(\mu_k, \Sigma_k), \qquad \sum_{k=1}^{K} \pi_k = 1.$$



Z_i **is the assignment**: the kind that produced car i. The model needs it, the table never contains it, and such a variable is called latent.

What the Three Parameters Mean

- π_k — **how common the kind is**: the share of all cars that are of kind k. If $\pi_2 = .26$, about a quarter of the market is kind 2.
- μ_k — **what it looks like**: the typical weight and mileage of kind k, a single point in the same plane as the data.
- Σ_k — **how much it varies**: how far cars of kind k spread around μ_k, and whether weight and mileage move together inside the kind.

Only μ_k has a counterpart in K-means. π_k and Σ_k are what turn membership into a share.

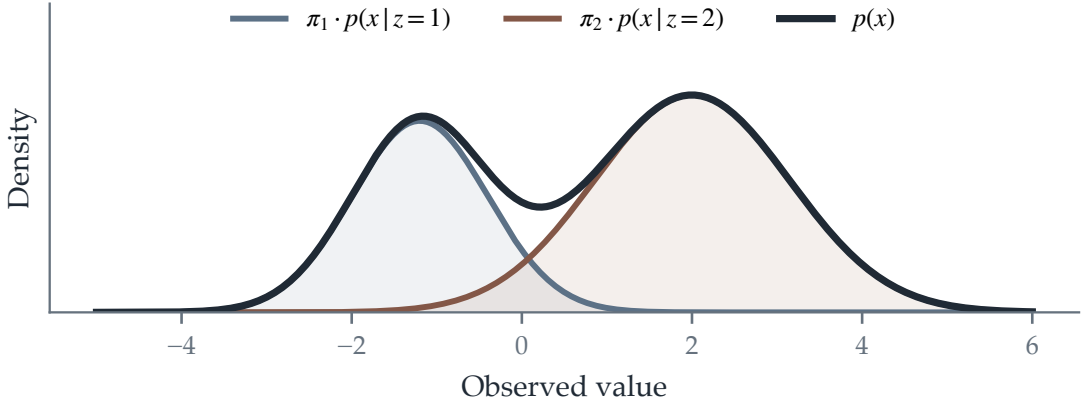
The Density That Follows

Averaging over the unseen assignment gives the density of a single car:

$$p(x) = \sum_{k=1}^{K} \Pr(Z = k) \cdot p(x \mid Z = k) = \sum_{k=1}^{K} \pi_k \cdot \mathcal{N}(x; \mu_k, \Sigma_k).$$

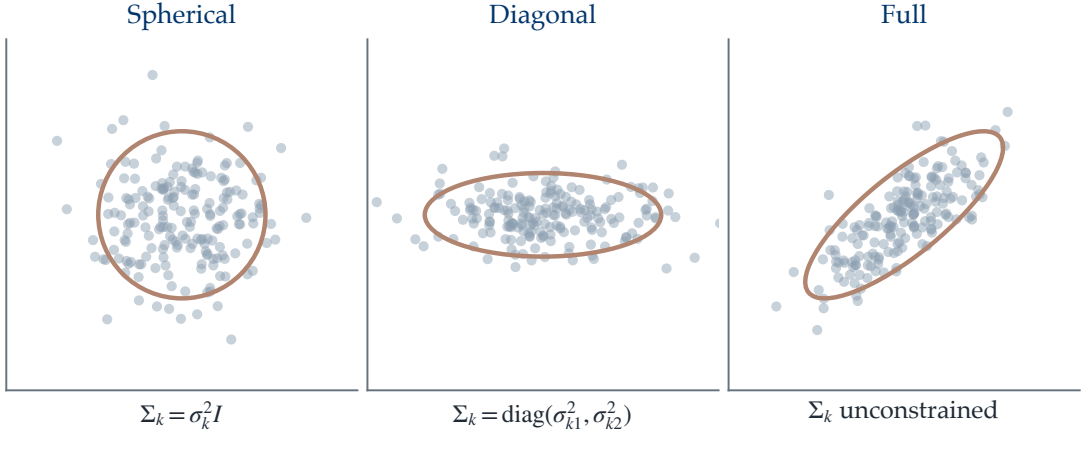
Every kind contributes to every car, weighted by how common that kind is.

A Mixture Adds Component Densities



Weighted component densities sum to the observed density.

Covariance Assumptions Define Component Geometry



Original Gaussian simulation, seed 26519; red curves are two-standard-deviation ellipses.

What a Probability Model Buys over a Partition

K-means returns labels and centers. The mixture returns a density.

- **How sure?** Membership is a probability: the Maverick is about half one kind and half another.
- **How common?** π_k estimates the size of each kind — a summary of the market, not only of the table.
- **Does the model fit this car?** $p_\theta(x)$ is small exactly when the model finds that car improbable.
- **Which model fits better?** Adding those up, $\sum_i \log p_\theta(x_i)$ grades the whole model on held-out cars: $K = 2$ against $K = 3$, spherical against full.

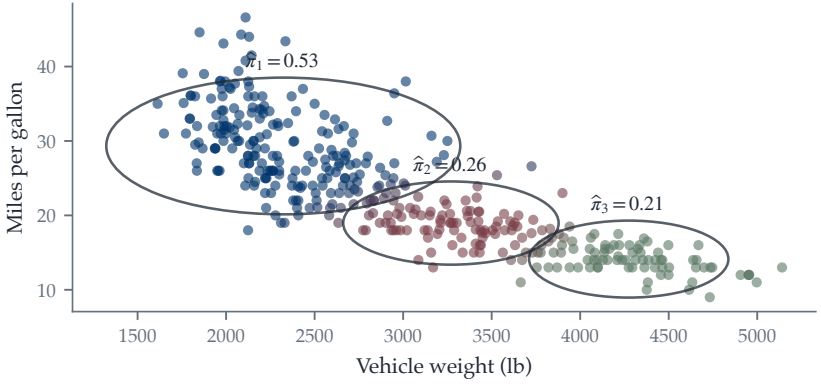
Where the Three Components Came From

We did not find three Gaussian populations in the table. We imposed them: the number K, the Gaussian shape, and independence across cars are assumptions, and none of them is checked by the data before we fit.

Two consequences to keep in view:

- Every conclusion inherits the assumption;
- Several Gaussians can approximate one non-Gaussian shape, so a fitted component need not be a real class. We come back to this at the end, when we finally look at the cylinder column.

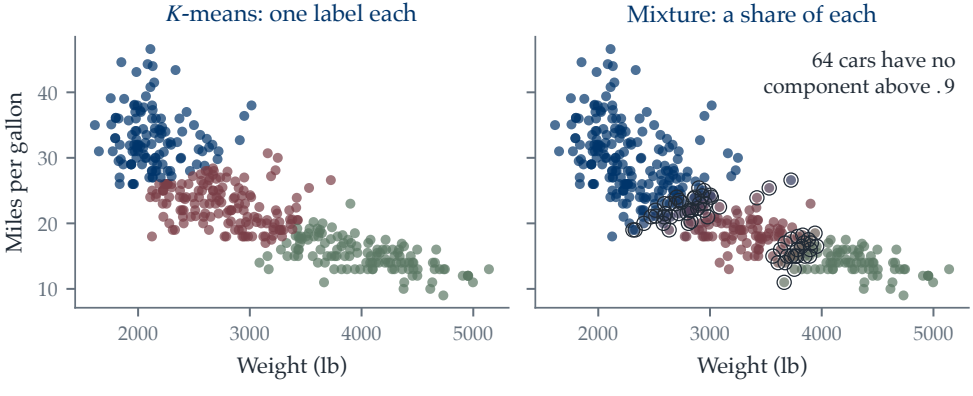
The Assumption, Fitted to the Cars



Three components: circles are two-standard-deviation contours, colour blends the three memberships, and $\hat{\pi}_k$ is each group's estimated prevalence. Where did these numbers come from? That question is the rest of the lecture.

ISLP Auto data; course spherical EM, seed 26519, best of twenty starts.

The Same Cars, Two Kinds of Answer



K-means states a label for every car with identical confidence. The mixture states a share, and 64 of 392 cars have no component above .9 — they sit exactly on the two seams the first picture drew straight through.

ISLP Auto data; *K*-means and spherical EM on the same standardized features, seed 26519.

Estimation by Maximum Likelihood

Write $\theta = (\pi_k, \mu_k, \Sigma_k)_{k=1}^{K}$ and choose it to make the observed cars as probable as possible:

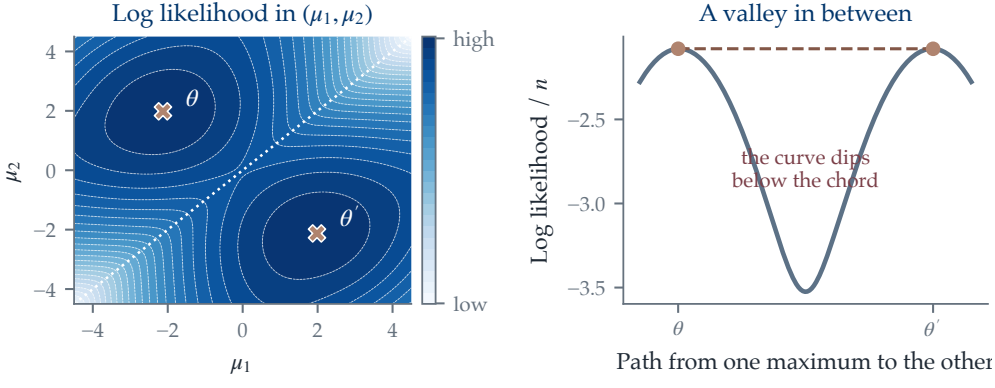
$$\hat{\theta} \in \arg\max_{\theta} \ \ell(\theta), \qquad \ell(\theta) = \sum_{i=1}^{n} \log p_\theta(x_i).$$

Substituting the mixture density,

$$\ell(\theta) = \sum_{i=1}^{n} \log \left[\sum_{k=1}^{K} \pi_k \cdot \mathcal{N}(x_i; \mu_k, \Sigma_k) \right].$$

The sum sits inside the logarithm. Nothing separates: every μ_k appears in every term, and ℓ is not concave.

Nonconcave Already with Two Means in One Dimension



Two components, unit variances, equal weights: only μ_1 and μ_2 are free. Relabelling maps θ to θ' without changing the likelihood, so there are two separated maxima — and the straight path between them runs through a valley. A concave function cannot do that.

Original simulation, seed 26519; 240 draws from an equal mixture of $\mathcal{N}(-2, 1)$ and $\mathcal{N}(2, 1)$.

The General Latent-Variable Model

A latent-variable model specifies a joint distribution over the hidden z and the observed x,

$$p_\theta(x, z) = p_\theta(z)\, p_\theta(x \mid z),$$

with both factors chosen to be simple. The mixture is the case $p_\theta(z = k) = \pi_k$ and $p_\theta(x \mid z = k) = \mathcal{N}(x; \mu_k, \Sigma_k)$.

Only x is recorded, so maximum likelihood must fit the marginal:

$$\ell(\theta) = \sum_{i=1}^{n} \log p_\theta(x_i) = \sum_{i=1}^{n} \log \sum_{z} p_\theta(x_i, z).$$

The logarithm never reaches the simple factors, and ℓ is nonconcave for the same reason as before.

Idea: Keep the Alternation, Soften the Assignment

K-means alternated

1. Send each point to its nearest center;
2. Recompute each center from the points sent to it.

We alternate

1. Guess the assignment: holding θ fixed, Bayes' rule turns each car into a share γ_{ik};
2. Refit π, μ, Σ, counting every car as partly in each kind.

Step 2 is tractable: with the shares held fixed, the logarithm sits on one Gaussian at a time again, and each component has a closed-form weighted fit.

What Remains

- **Next — the assignment.** Bayes' rule turns the current parameters into $\gamma_{ik} = \Pr(Z_i = k \mid x_i)$.
- **Then — the parameters.** Weighted means, weighted covariances, and π_k as a fractional count.
- **Then — why it works.** The two steps are climbing a lower bound on $\ell(\theta)$, the evidence lower bound, which is why the likelihood never decreases.

Together the two steps are the expectation–maximization algorithm.

Outline

The Model Already Supplies a Prior and a Likelihood

Fix the parameters θ at their current values. The mixture then states two things about a single car.

Prior — before seeing x_i

$$\Pr(Z_i = k) = \pi_k$$

how common kind k is.

Likelihood — if kind k produced it

$$p(x_i \mid Z_i = k) = \mathcal{N}(x_i; \mu_k, \Sigma_k)$$

how plausible x_i looks under kind k.

What we want is the reverse question: given that we observed x_i, which kind produced it?

Computing the Posterior

Bayes' rule combines them — posterior $\propto$ prior $\times$ likelihood, normalized over the K kinds:

$$\gamma_{ik} \;=\; \Pr(Z_i = k \mid x_i) \;=\; \frac{\Pr(Z_i = k) \cdot p(x_i \mid Z_i = k)}{\sum_{\ell=1}^{K} \Pr(Z_i = \ell) \cdot p(x_i \mid Z_i = \ell)},$$

Computing the Posterior

Bayes' rule combines them — posterior $\propto$ prior $\times$ likelihood, normalized over the K kinds:

$$\gamma_{ik} \;=\; \Pr(Z_i = k \mid x_i) \;=\; \frac{\Pr(Z_i = k) \cdot p(x_i \mid Z_i = k)}{\sum_{\ell=1}^{K} \Pr(Z_i = \ell) \cdot p(x_i \mid Z_i = \ell)},$$

which for the Gaussian mixture is

$$\gamma_{ik} = \frac{\pi_k \cdot \mathcal{N}(x_i; \mu_k, \Sigma_k)}{\sum_{\ell=1}^{K} \pi_\ell \cdot \mathcal{N}(x_i; \mu_\ell, \Sigma_\ell)}.$$

Computing the Posterior

Bayes' rule combines them — posterior $\propto$ prior $\times$ likelihood, normalized over the K kinds:

$$\gamma_{ik} \;=\; \Pr(Z_i = k \mid x_i) \;=\; \frac{\Pr(Z_i = k) \cdot p(x_i \mid Z_i = k)}{\sum_{\ell=1}^{K} \Pr(Z_i = \ell) \cdot p(x_i \mid Z_i = \ell)},$$

which for the Gaussian mixture is

$$\gamma_{ik} = \frac{\pi_k \cdot \mathcal{N}(x_i; \mu_k, \Sigma_k)}{\sum_{\ell=1}^{K} \pi_\ell \cdot \mathcal{N}(x_i; \mu_\ell, \Sigma_\ell)}.$$

The denominator is $p_\theta(x_i)$, the mixture density from the last section.

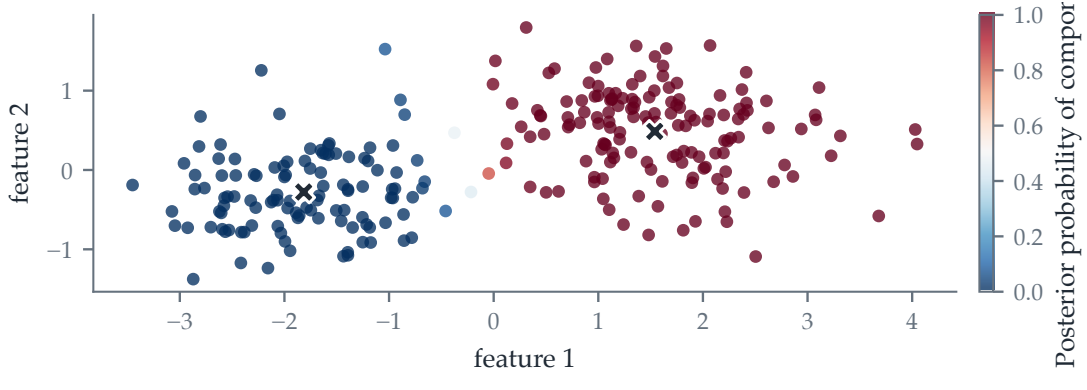
What γ_{ik} Means

γ_{ik} is a **posterior probability**: given the evidence x_i, how likely it is that car i came from kind k.

- Before seeing the car, kind k had probability π_k. After seeing its weight and mileage, that becomes γ_{ik}: the evidence has reweighted the kinds, and $\sum_k \gamma_{ik} = 1$ still holds.
- It is conditional on θ. Change the parameters and every posterior changes — γ_{ik} is the model's current belief, not a fact about the car.
- A hard label is the extreme case $\gamma_{ik} \in \{0, 1\}$, and K-means reports nothing else.

Many textbooks call γ_{ik} the **responsibility** of component k for observation i.

The Posterior Preserves Ambiguity



Original course simulation, seed 26519; colour is the posterior probability of component 2.

One Posterior Calculation

Suppose equal priors, unit variances, means 0 and 3, and observation $x = 1$. The common Gaussian constant cancels:

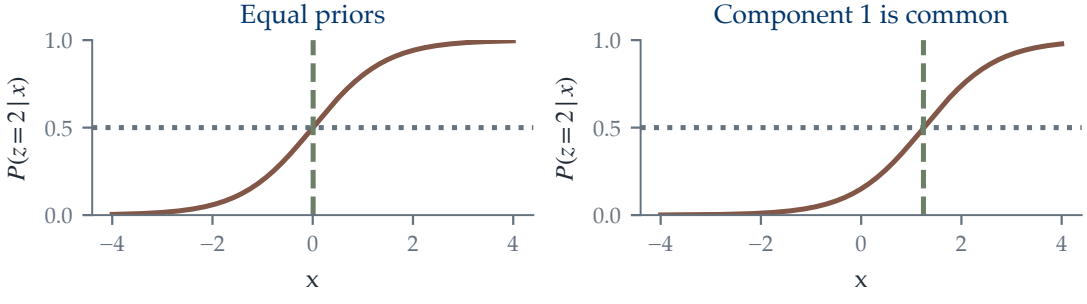
$$w_1 = \tfrac{1}{2} e^{-(1-0)^2/2} = \tfrac{1}{2} e^{-1/2}, \qquad w_2 = \tfrac{1}{2} e^{-(1-3)^2/2} = \tfrac{1}{2} e^{-2}.$$

Therefore

$$\gamma_1 = \frac{w_1}{w_1 + w_2} = \frac{e^{-1/2}}{e^{-1/2} + e^{-2}} \approx .818, \qquad \gamma_2 \approx .182.$$

A hard nearest-mean rule would discard this uncertainty.

Component Priors Shift Posterior Membership



Only the component prior weights differ.

Mean Distance Is Only One Factor

Taking logarithms of the posterior numerator gives

$$\log \pi_k - \frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (x_i - \mu_k)^\top \Sigma_k^{-1} (x_i - \mu_k) + \text{constant}.$$

- π_k favors common components;
- $|\Sigma_k|$ controls density concentration;
- Mahalanobis distance uses component-specific shape.

Thus the posterior is not decided by nearest Euclidean mean alone.

Stable Log-Sum-Exp Calculation

derive

Define $a_{ik} = \log \pi_k + \log \mathcal{N}(x_i; \mu_k, \Sigma_k)$ and $m_i = \max_k a_{ik}$. Then

$$\begin{aligned} \log \sum_k e^{a_{ik}} &= \log \left[e^{m_i} \sum_k e^{a_{ik} - m_i} \right] \\ &= m_i + \log \sum_k e^{a_{ik} - m_i}. \end{aligned}$$

Every exponent $a_{ik} - m_i \leq 0$, preventing large exponentials and reducing underflow before normalization.

Next: the Other Half of the Loop

With θ in hand we can compute every γ_{ik}. The remaining question is the reverse one:

given the posteriors, how do we re-estimate π_k, μ_k, and Σ_k?

Next we write the likelihood we would have had if the assignments were observed, and shows that replacing each label by its γ_{ik} makes that fit a weighted version of the ordinary Gaussian estimates.

Outline

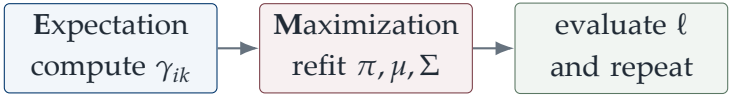
Complete-Data Log Likelihood

If labels were observed, write $z_{ik} = \mathbf{1}\{Z_i = k\}$. Then

$$\ell_c(\theta) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \left[\log \pi_k + \log \mathcal{N}(x_i; \mu_k, \Sigma_k) \right].$$

Conditional on z, parameters separate by component. EM replaces the missing indicator by its posterior expectation and repeats.

One Cycle: Expectation, then Maximization



The two letters name what each half does. The **E-step** takes the expectation of $\ell_c(\theta)$ over the unknown assignments, using the posterior at the current parameters. The **M-step** maximizes that expected likelihood over θ.

The E-Step: an Expectation over the Unknown Z

We cannot evaluate $\ell_c(\theta)$, because z_{ik} is not observed. Replace each indicator by its expected value under the current posterior:

$$\mathbb{E}[z_{ik} \mid x_i, \theta^{(t)}] = \Pr(Z_i = k \mid x_i, \theta^{(t)}) = \gamma_{ik}^{(t)}.$$

Carrying that through ℓ_c gives the objective the M-step will maximize:

$$Q(\theta \mid \theta^{(t)}) = \mathbb{E}_{Z|X,\theta^{(t)}}[\ell_c(\theta)] = \sum_{i,k} \gamma_{ik}^{(t)} [\log \pi_k + \log \mathcal{N}(x_i; \mu_k, \Sigma_k)].$$

Every fractional assignment is kept; no hard label is imputed.

The M-Step: Maximize Q over θ

derive

Define the effective count $N_k = \sum_i \gamma_{ik}$. Maximizing $\sum_k N_k \log \pi_k$ subject to $\sum_k \pi_k = 1$ gives

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

For fixed covariance, differentiating the weighted quadratic yields

$$0 = \sum_i \gamma_{ik}(x_i - \mu_k) \quad \Longrightarrow \quad \boxed{\mu_k^{\text{new}} = \frac{1}{N_k} \sum_i \gamma_{ik} x_i}.$$

N_k plays the role of a fractional sample size.

Covariance Update

The weighted covariance is

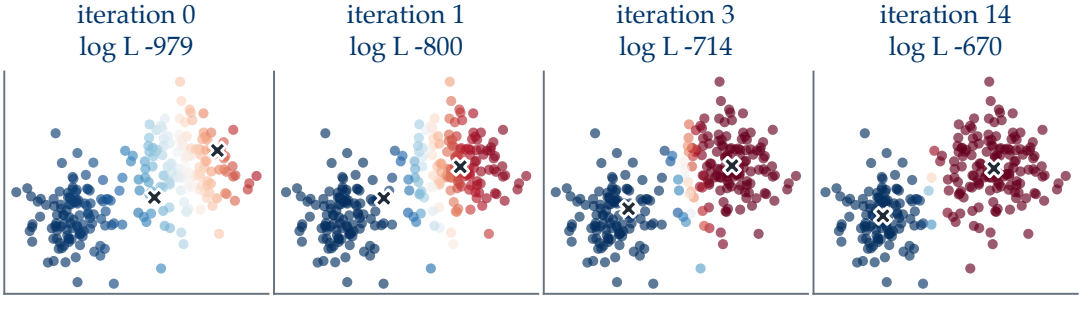
$$\Sigma_k^{\text{new}} = \frac{1}{N_k} \sum_{i=1}^{n} \gamma_{ik} (x_i - \mu_k^{\text{new}})(x_i - \mu_k^{\text{new}})^\top.$$

A diagonal family keeps only diagonal entries. For spherical covariance in d dimensions,

$$(\sigma_k^2)^{\text{new}} = \frac{1}{dN_k} \sum_i \gamma_{ik} \|x_i - \mu_k^{\text{new}}\|_2^2.$$

Covariance constraints trade geometric flexibility for statistical stability.

Posteriors and Means Co-Evolve



Original spherical-mixture simulation; colour is the posterior for the second component and crosses are means.

One M-Step Calculation

For $x = (0, 1, 4)$ and component-1 posteriors $(.9, .8, .1)$,

$$N_1 = .9 + .8 + .1 = 1.8, \qquad \pi_1^{\text{new}} = 1.8/3 = .6,$$

$$\mu_1^{\text{new}} = \frac{.9(0) + .8(1) + .1(4)}{1.8} = \frac{1.2}{1.8} = .667.$$

Complementary posteriors give $N_2 = 1.2$ and $\pi_2^{\text{new}} = .4$. The component means are weighted, not based on thresholded labels.

Outline

A Lower Bound on the Log Likelihood

Take any distribution $q(z)$ over the hidden variable, and multiply and divide by it inside the same sum as before:

$$\log p_\theta(x) = \log \sum_z p_\theta(x,z) = \log \sum_z q(z) \cdot \frac{p_\theta(x,z)}{q(z)}$$

$$\geq \sum_z q(z) \cdot \log \frac{p_\theta(x,z)}{q(z)} \equiv \mathcal{L}(q,\theta).$$

$$\boxed{\log p_\theta(x) \geq \mathcal{L}(q,\theta) \quad \text{for every } q}$$

$\mathcal{L}$ is the **evidence lower bound**, or ELBO. The logarithm has moved inside the sum, so it now sits on the joint $p_\theta(x,z)$, which factorizes.

Source: The inequality is Jensen's; derived in Appendix A.1.

Inside the Bound, the Joint Splits

The joint is prior times likelihood, $p_\theta(x, z) = p_\theta(z) \cdot p_\theta(x \mid z)$, and the logarithm now reaches both factors:

$$\mathcal{L}(q, \theta) = \sum_z q(z)[\log p_\theta(z) + \log p_\theta(x \mid z)] \; - \; \sum_z q(z) \log q(z).$$

Inside the Bound, the Joint Splits

The joint is prior times likelihood, $p_\theta(x, z) = p_\theta(z) \cdot p_\theta(x \mid z)$, and the logarithm now reaches both factors:

$$\mathcal{L}(q, \theta) = \sum_z q(z)[\log p_\theta(z) + \log p_\theta(x \mid z)] \; - \; \sum_z q(z) \log q(z).$$

The second sum has no θ in it, so it is a constant as far as the M-step is concerned:

$$\arg\max_\theta \; \mathcal{L}(q, \theta) = \arg\max_\theta \sum_z q(z) \log p_\theta(x, z) = \arg\max_\theta \; \mathbb{E}_q[\log p_\theta(x, z)].$$

For the mixture this is the Q already derived, with $q(k)$ in place of γ_{ik}: maximizing the bound over θ is the weighted fit already derived.

The Same Objects, in the Gaussian Mixture

	general latent-variable model	Gaussian mixture, car i
prior	$p_\theta(z)$	$\Pr(Z_i = k) = \pi_k$
likelihood	$p_\theta(x \mid z)$	$\mathcal{N}(x_i; \mu_k, \Sigma_k)$
joint	$p_\theta(x, z)$	$\pi_k \cdot \mathcal{N}(x_i; \mu_k, \Sigma_k)$
marginal	$p_\theta(x) = \sum_z p_\theta(x, z)$	$\sum_k \pi_k \cdot \mathcal{N}(x_i; \mu_k, \Sigma_k)$
posterior	$p_\theta(z \mid x)$	γ_{ik}
free distribution	$q(z)$, any distribution	$(q_{i1}, \dots, q_{iK})$ with $\sum_k q_{ik} = 1$

The bound holds for every choice of q_i. The next slide picks the one that makes it tight.

EM in the Abstract Form: Alternating Optimization on ELBO

$\mathcal{L}(q, \theta)$ has two arguments, so maximize it in one and then the other:

$$q^{(t+1)} = \arg\max_{q} \ \mathcal{L}(q, \theta^{(t)}), \qquad \theta^{(t+1)} = \arg\max_{\theta} \ \mathcal{L}(q^{(t+1)}, \theta).$$

The second is the weighted maximum-likelihood fit already derived. The first looks harder — it optimizes over distributions, not numbers.

EM in the Abstract Form: Alternating Optimization on ELBO

$\mathcal{L}(q, \theta)$ has two arguments, so maximize it in one and then the other:

$$q^{(t+1)} = \arg\max_{q} \; \mathcal{L}(q, \theta^{(t)}), \qquad \theta^{(t+1)} = \arg\max_{\theta} \; \mathcal{L}(q^{(t+1)}, \theta).$$

The second is the weighted maximum-likelihood fit already derived. The first looks harder — it optimizes over distributions, not numbers.

It has a closed-form answer.

The Best q Is the Posterior

Write $D_{\mathrm{KL}}(q \,\|\, p) = \sum_z q(z) \log \frac{q(z)}{p(z)}$ for the **Kullback–Leibler divergence**: it is never negative and is zero exactly when $q = p$. Then, for every q,

$$\boxed{\log p_\theta(x) = \mathcal{L}(q, \theta) + D_{\mathrm{KL}}(q \,\|\, p_\theta(z \mid x))}$$

The left side does not involve q, so maximizing $\mathcal{L}$ over q is the same as driving that gap to zero:

$$q^{(t+1)}(z) = p_{\theta^{(t)}}(z \mid x), \qquad \mathcal{L}(q^{(t+1)}, \theta^{(t)}) = \log p_{\theta^{(t)}}(x).$$

That is the E-step, and for the mixture this posterior is $(\gamma_{i1}, \dots, \gamma_{iK})$.

Source: Identity derived in Appendix A.2.

The M-Step Raises the Bound

Hold $q^{(t+1)}$ fixed and choose $\theta^{(t+1)}$ to increase the ELBO:

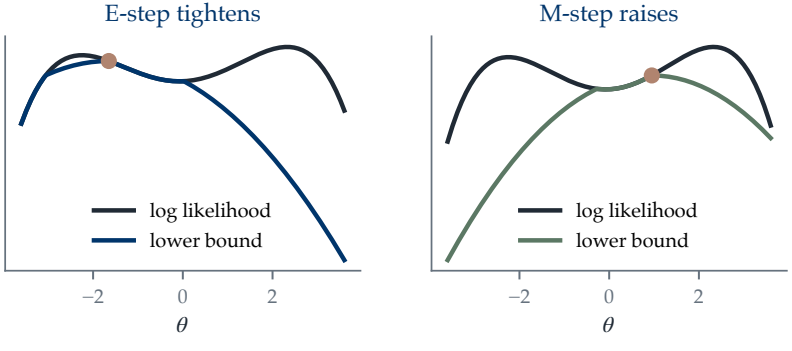
$$\mathcal{L}(q^{(t+1)}, \theta^{(t+1)}) \geq \mathcal{L}(q^{(t+1)}, \theta^{(t)}).$$

Since the log likelihood is always at least its ELBO,

$$\log p_{\theta^{(t+1)}}(x) \geq \mathcal{L}(q^{(t+1)}, \theta^{(t+1)}).$$

Combining these inequalities proves observed-likelihood monotonicity.

Tighten the Bound, Then Raise It



Source: Original pedagogical schematic; not a fitted-mixture trace.

Why EM Cannot Go Downhill

Chaining the two steps gives

$$\boxed{\log p_{\theta^{(t+1)}}(x) \;\geq\; \log p_{\theta^{(t)}}(x)}$$

for every iteration: the observed likelihood never decreases.

Why EM Cannot Go Downhill

Chaining the two steps gives

$$\boxed{\log p_{\theta^{(t+1)}}(x) \;\geq\; \log p_{\theta^{(t)}}(x)}$$

for every iteration: the observed likelihood never decreases. The E-step makes the bound touch the likelihood, so raising the bound in the M-step must carry the likelihood with it.

Why EM Cannot Go Downhill

Chaining the two steps gives

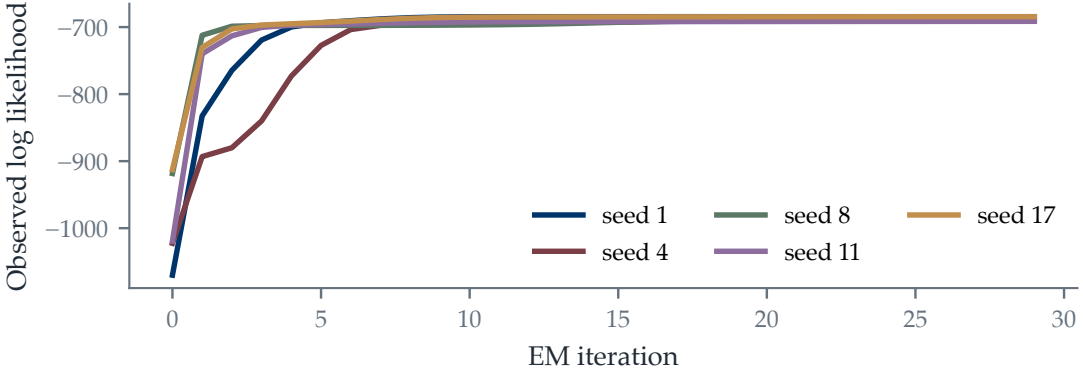
$$\log p_{\theta^{(t+1)}}(x) \;\geq\; \log p_{\theta^{(t)}}(x)$$

for every iteration: the observed likelihood never decreases. The E-step makes the bound touch the likelihood, so raising the bound in the M-step must carry the likelihood with it.

Monotone is not optimal. It guarantees no global maximum, no unique parameters, and no meaningful interpretation of a component — only that each step is not a step backwards.

Source: Three-line chain in Appendix A.4.

Monotone Restarts Can End Differently



Every run climbs, and they do not all arrive at the same place. Compare runs by their final likelihood, never by component number: relabelling the components leaves the likelihood unchanged, so “component 2” in one run has nothing to do with “component 2” in another.

Original three-component simulation with five initializations.

Outline

Back to the Answer Key

The three fitted components, in the original units, beside the withheld cylinder column:

component	$\widehat{\pi}_k$	weight	mpg	4 cyl	6 cyl	8 cyl
1	.53	2,323	29	190	13	1
2	.26	3,270	19	9	65	22
3	.21	4,274	14	0	5	80

The model never saw this column, yet the largest-posterior assignment reproduces it for **335 of 385** cars; the 50 disagreements sit in the overlap.

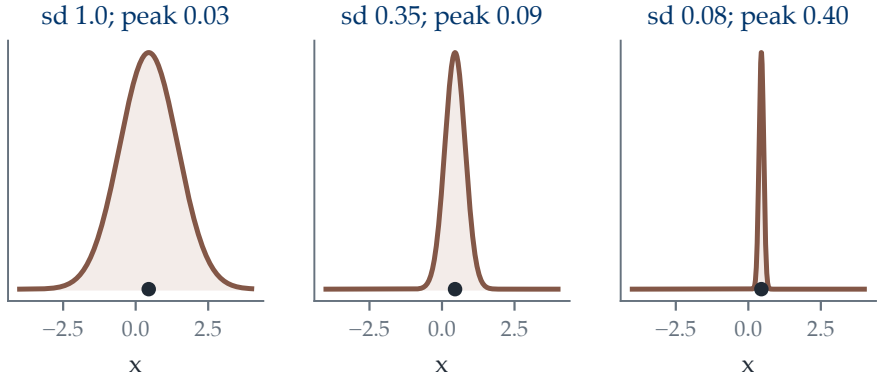
Source: ISLP Auto data; seven 3- or 5-cylinder cars omitted.

Here We Could Check. Usually We Cannot

On this data the components track a real engineering distinction rather than an arbitrary cut through a smooth cloud — and we know that only because a column existed to compare against.

Most problems have no such column. The remaining slides are the diagnostics that work without one: what goes wrong, and how it shows itself in the fit rather than in an answer key.

A Component Can Collapse onto One Observation



This is the degenerate case where a component ends up with effectively one data point — $N_k \to 1$ — so its mean sits exactly on that point, the weighted spread around it is zero, and its density there grows without bound.

Original one-dimensional calculation; component weight is .08.

Why Gaussian-Mixture Likelihood Can Diverge

derive

For one component centered exactly at observation x_i, a d-dimensional spherical density is

$$\mathcal{N}(x_i; x_i, \sigma^2 I) = (2\pi\sigma^2)^{-d/2}.$$

As $\sigma^2 \to 0$,

$$(2\pi\sigma^2)^{-d/2} \to \infty, \qquad \log p_\theta(x_i) \to \infty.$$

Collapse is a property of the unregularized likelihood, not only floating-point arithmetic.

Preventing Degenerate Components

Change the fitted problem

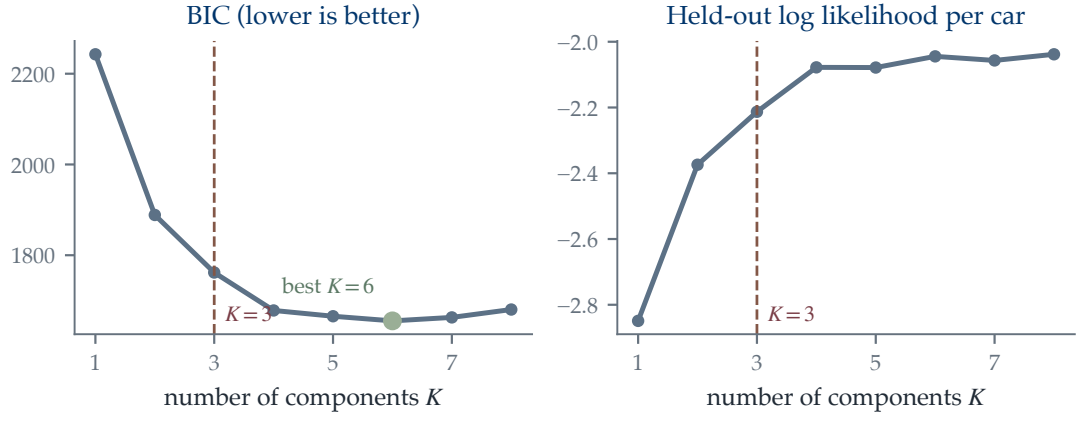
variance floors, tied covariance, penalties, or priors prevent arbitrarily sharp components

Inspect the result

multiple starts, effective counts N_k, held-out likelihood, and density plots reveal fragile fits

Regularization should be declared as part of the model, not hidden as a numerical patch.

Choosing K on the Cars



Both curves improve steeply to $K = 3$ or 4 and then flatten — and BIC's actual minimum is at $K = 6$. Past the elbow the extra components are tiling a curved band, not finding new kinds of car.

ISLP Auto data; course spherical EM, best of ten starts per K; five-fold split, seed 26519.

A Rule of Thumb for K

- Fit $K = 1, 2, \ldots$ and plot both curves. Take the elbow, the smallest K after which the gain is small — not the strict optimum.
- Check the smallest effective count $N_k = \sum_i \gamma_{ik}$: here it falls from 82 at $K = 3$ to 14 at $K = 8$. A component holding a handful of points is an artifact, and a collapse waiting to happen.
- **BIC** $= -2\ell + p \log n$ charges for parameters ($p = 4K - 1$ here; see Appendix A.3); **held-out likelihood** charges nothing but needs a split.

A good density and a meaningful set of kinds are different goals, and they do not always want the same K.

From Exact EM to Variational Inference

Exact E-step

compute $p_\theta(z \mid x)$ in closed form and make the ELBO tight

Approximate inference

choose $q_\phi(z \mid x)$ from a tractable family and optimize an approximate ELBO

From Exact EM to Variational Inference

Exact E-step	Approximate inference
compute $p_\theta(z \mid x)$ in closed form and make the ELBO tight	choose $q_\phi(z \mid x)$ from a tractable family and optimize an approximate ELBO

Two things break when the model is a neural network: $p_\theta(z \mid x)$ has no closed form, and recomputing it per observation is too slow. A **variational autoencoder** answers both by learning the E-step — one encoder network $q_\phi(z \mid x)$, shared across all observations, trained jointly with θ by maximizing the same ELBO.

The objective in Lecture 17 is the bound on this slide. Only the way the E-step is obtained changes.

Summary: Gaussian Mixtures and EM

$$\gamma_{ik} = \frac{\pi_k p_k(x_i)}{\sum_\ell \pi_\ell p_\ell(x_i)}, \qquad \mu_k^{\text{new}} = \frac{\sum_i \gamma_{ik} x_i}{\sum_i \gamma_{ik}}.$$

1. Mixtures model a density through latent component labels and soft membership.
2. EM alternates posterior inference with weighted maximum likelihood.
3. Priors and covariance geometry matter in addition to mean distance.

Summary: Bounds and Failure Modes

$$\log p_\theta(x) = \mathcal{L}(q, \theta) + D_{\mathrm{KL}}(q \,\|\, p_\theta(z \mid x)).$$

1. The E-step tightens the ELBO and the M-step raises it, so likelihood does not decrease.
2. Local maxima and covariance collapse require explicit safeguards; component numbers are arbitrary and carry none.
3. A component is model-conditional and need not be a real or causal class.

Next Lecture

EM recomputes the posterior separately for every observation, and it needs that posterior in closed form. Neither survives once the model that generates x from z is a neural network.

Lecture 17 — Autoencoders and Variational Autoencoders replaces the E-step by a single encoder network that outputs an approximate posterior for any input, and trains it against the bound derived today.

Reading for this lecture: ISLP Chapter 12; DL Chapter 13.

A.1 The Jensen Lower Bound

derive

Insert any distribution $q(z)$ whose support covers the joint model, then apply Jensen's inequality to the concave logarithm:

$$\log p_\theta(x) = \log \sum_z p_\theta(x,z) = \log \sum_z q(z)\,\frac{p_\theta(x,z)}{q(z)}$$
$$\geq \sum_z q(z) \log \frac{p_\theta(x,z)}{q(z)} \equiv \mathcal{L}(q,\theta).$$

Equality holds when $p_\theta(x,z)/q(z)$ is constant in z, which forces $q(z) = p_\theta(z \mid x)$.

A.2 The Bound Plus the KL Gap

derive

Splitting the joint as $p_\theta(x, z) = p_\theta(z \mid x)\, p_\theta(x)$ inside the bound gives

$$\begin{aligned}\mathcal{L}(q, \theta) &= \sum_z q(z) \log \frac{p_\theta(z \mid x)\, p_\theta(x)}{q(z)} \\ &= \log p_\theta(x) - \sum_z q(z) \log \frac{q(z)}{p_\theta(z \mid x)} = \log p_\theta(x) - D_{\mathrm{KL}}(q \,\|\, p_\theta(z \mid x)).\end{aligned}$$

Because $D_{\mathrm{KL}} \geq 0$ with equality only at $q = p_\theta(z \mid x)$, the bound is tight exactly at the posterior, which is the E-step.

A.3 The Bayesian Information Criterion

Rather than hold data out, **BIC** charges the training fit for the parameters it used:

$$\mathrm{BIC} = -2\,\ell(\hat{\theta}) + p \log n, \qquad \text{smaller is better,}$$

where p counts free parameters and n counts observations (here, 392 cars).

- A spherical mixture in d dimensions has $p = K(d+1) + (K-1)$: a mean and a variance per component, plus $K-1$ weights. At $d=2$, $p = 4K-1$.
- Each parameter costs $\log n$, so more data means an extra component must earn more before it is worth keeping.
- It approximates a Bayesian model comparison under conditions mixtures do not strictly satisfy: a guide, not a test.

Source: Read ISLP Section 6.1.3, “Choosing the Optimal Model” (pp. 235–239).

A.4 The Monotonicity Chain

The complete argument is

$$\begin{aligned}\log p_{\theta^{(t+1)}}(x) &\geq \mathcal{L}(q^{(t+1)}, \theta^{(t+1)}) \\ &\geq \mathcal{L}(q^{(t+1)}, \theta^{(t)}) \\ &= \log p_{\theta^{(t)}}(x).\end{aligned}$$

This guarantees nondecreasing observed likelihood. It does not guarantee a global maximum, unique parameters, or a scientifically meaningful component interpretation.