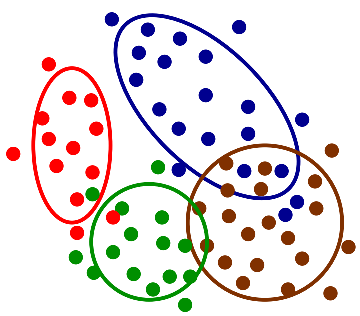
Latent variable models

Mixture models

Unsupervised classification

Unsupervised classification

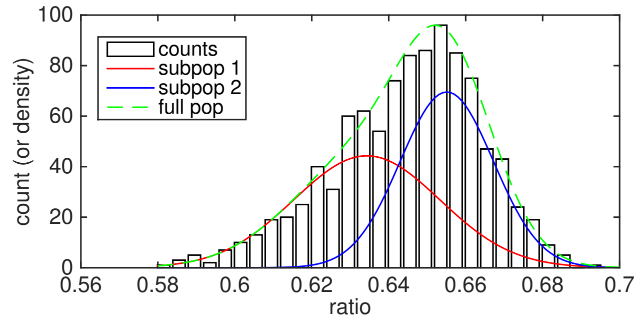
- **Input**: $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$, number of classes $K \in \mathbb{N}$.
- **Output**: function $f \colon \mathbb{R}^d \to \{1, 2, \ldots, K\}$.
  (Notation: $[K] := \{1, 2, \ldots, K\}$.)
- **Typical semantics**: hidden subpopulation structure.



Example: Pearson's crabs (1894)

Data: ratio of forehead-width to body-length for 1000 crabs.

Maybe the sample is comprised of two different sub-species of crab?



Gaussian mixture model

Gaussian mixture model: statistical model $\mathcal{P} = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ on $\mathcal{X} \times [K]$, where $(\boldsymbol{X}, Y) \sim P_{\boldsymbol{\theta}}$ means that

$$Y \sim (\pi_1, \dots, \pi_K) \quad \text{(discrete distribution over } [K]\text{; } P_{\boldsymbol{\theta}}(Y=j) = \pi_j\text{)}$$
$$\boldsymbol{X} \mid Y = j \sim \mathrm{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \quad \text{(Gaussian with mean } \boldsymbol{\mu}_j \text{ and covariance } \boldsymbol{\Sigma}_j\text{)}$$

Parameter space $\boldsymbol{\Theta}$ comprises all $\boldsymbol{\theta} = (\pi_1, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \dots, \pi_K, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K)$ where $\boldsymbol{\mu}_j \in \mathbb{R}^d$, $\boldsymbol{\Sigma}_j \succeq \mathbf{0}$ (positive definite $d \times d$ matrix), $\pi_j \in [0,1]$, and $\sum_{j=1}^K \pi_j = 1$.

Looks familiar?

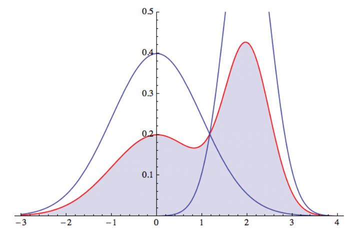
Even though this model is for data $(\boldsymbol{x}, y) \in \mathbb{R}^d \times [K]$, we declare only the $\boldsymbol{x}$ part to be **observable**, and declare the y part to be **hidden** (or **latent**).

Models of this sort are called **mixture models**;
this one in particular is called the **Gaussian mixture model**.

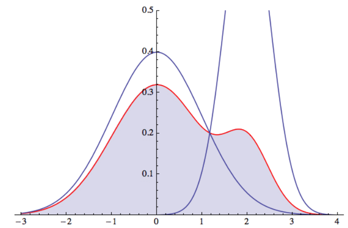
$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{j=1}^K \pi_j \cdot (2\pi)^{-d/2} \sqrt{\det(\boldsymbol{\Sigma}_j^{-1})} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_j)^\top \boldsymbol{\Sigma}_j^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_j)\right)$$

Mixing weights $\boldsymbol{\pi}$; **mixture components** $\mathrm{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \dots, \mathrm{N}(\boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K)$.

Gaussian mixtures in $\mathbb{R}^1$

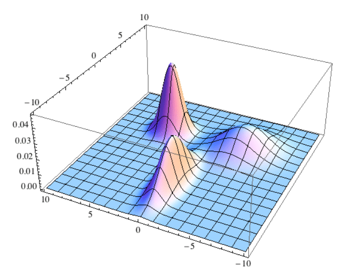


$$\frac{1}{2}\mathrm{N}(0,1) + \frac{1}{2}\mathrm{N}(2,1/4)$$

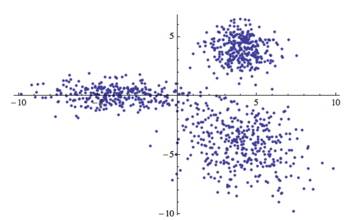


$$\frac{4}{5}\mathrm{N}(0,1) + \frac{1}{5}\mathrm{N}(2,1/4)$$

Gaussian mixtures in $\mathbb{R}^2$



Plot of the mixture density.



An iid sample of size 1000.

Soft assignments

Suppose you have the parameters $\boldsymbol{\theta} = (\pi_1, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \dots, \pi_K, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K)$ of a Gaussian mixture distribution, and further that $(\boldsymbol{X}, Y) \sim P_{\boldsymbol{\theta}}$.

Assignment variables $\boldsymbol{\Phi} = (\Phi_1, \Phi_2, \dots, \Phi_K) \in \{0,1\}^K$ (as in K-means):

$$\Phi_j := \mathbb{1}\{Y = j\}.$$

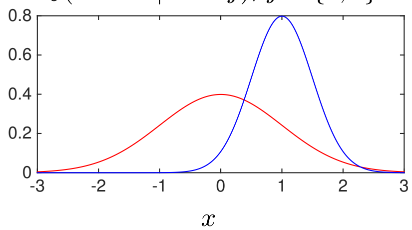
Soft assignment of a data point $\boldsymbol{x} \in \mathbb{R}^d$ to component $j \in [K]$:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta}}[\Phi_j \mid \boldsymbol{X} = \boldsymbol{x}] &= P_{\boldsymbol{\theta}}(Y = j \mid \boldsymbol{X} = \boldsymbol{x}) \\
&= \frac{P_{\boldsymbol{\theta}}(Y=j) \cdot p_{\boldsymbol{\theta}}(\boldsymbol{x} \mid Y = j)}{p_{\boldsymbol{\theta}}(\boldsymbol{x})} \\
&= \frac{\pi_j \cdot \sqrt{\det(\boldsymbol{\Sigma}_j^{-1})} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_j)^\top \boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_j)\right)}{\sum_{j'=1}^K \pi_{j'} \cdot \sqrt{\det(\boldsymbol{\Sigma}_{j'}^{-1})} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_{j'})^\top \boldsymbol{\Sigma}_{j'}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_{j'})\right)}.
\end{aligned}$$

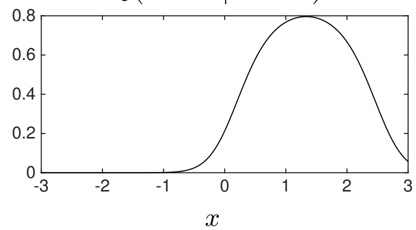
Soft clustering

Example: a Gaussian mixture distribution with $k = 2$ in $\mathbb{R}^1$.

$P_{\boldsymbol{\theta}}(X = x \mid Y = j)$, $j \in \{1, 2\}$.



$P_{\boldsymbol{\theta}}(Y = 1 \mid X = x)$



$$P_{\boldsymbol{\theta}}(Y = 1 \mid X = x) \;=\; \frac{\pi_1 \cdot \frac{1}{\sigma_1} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right)}{\pi_1 \cdot \frac{1}{\sigma_1} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right) + \pi_2 \cdot \frac{1}{\sigma_2} \exp\left(-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right)}.$$

Parameter estimation for Gaussian mixtures

Maximum likelihood estimation of $\boldsymbol{\theta} = (\pi_1, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \dots, \pi_K, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K)$ given data $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n$ (regarded as an i.i.d. sample).

$$\begin{aligned} \hat{\boldsymbol{\theta}} &:= \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sum_{i=1}^n \ln p_{\boldsymbol{\theta}}(\boldsymbol{x}_i) \\ &= \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sum_{i=1}^n \ln \left\{ \sum_{j=1}^K \pi_j \cdot \sqrt{\det(\boldsymbol{\Sigma}_j^{-1})} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_j)^\top \boldsymbol{\Sigma}_j^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_j)\right) \right\} \end{aligned}$$

Argh! The $\ln \left\{ \sum_{j=1}^K \cdots \right\}$ does not simplify nicely!

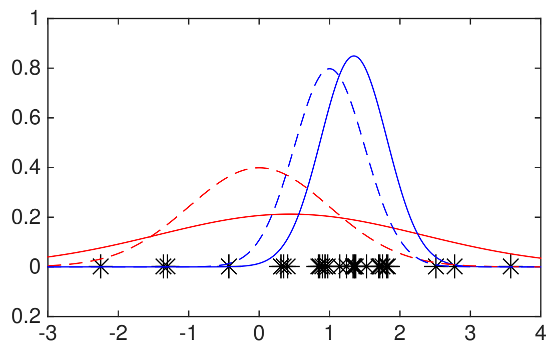
MLE for Gaussian mixtures: not a convex optimization problem.

$$\arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \quad -\sum_{i=1}^n \ln \left\{ \sum_{j=1}^K \pi_j \cdot \sqrt{\det(\boldsymbol{\Sigma}_j^{-1})} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_j)^\top \boldsymbol{\Sigma}_j^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_j)\right) \right\}$$

Local optima not guaranteed to be global optima;
could be arbitrarily far from / worse than the MLE.

Local optimization

- For the purpose of modeling the density of $\boldsymbol{X}$, **a "good enough" local maximizer could be sufficient**.
- If the data are actually generated by a Gaussian mixture distribution $P_{\boldsymbol{\theta}_\star}$, **then $\theta_\star$ may be close to some local maximizer of the likelihood**.



Methods like gradient ascent would work, **but there's a much nicer local optimization method for this case: the E-M algorithm.**

Expectation-Maximization for Gaussian mixtures

Motivating derivation

Suppose we had *softly labeled* data $\{(\boldsymbol{x}_i, \boldsymbol{w}_i)\}_{i=1}^n$ from $\mathbb{R}^d \times [0,1]^K$.
(Each $\boldsymbol{w}_i = (w_{i,1}, w_{i,2}, \dots, w_{i,K})$ is a probability distribution on $[K]$.)

The "complete log-likelihood" of $\boldsymbol{\theta} = (\pi_1, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \dots, \pi_K, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K)$ is

$$\sum_{i=1}^n \sum_{j=1}^K w_{i,j} \ln\left\{ \pi_j \cdot \sqrt{\det(\boldsymbol{\Sigma}_j^{-1})} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_j)^\top \boldsymbol{\Sigma}_j^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_j) \right) \right\}$$
$$= \sum_{i=1}^n \sum_{j=1}^K w_{i,j} \left(\ln \pi_j + \frac{1}{2} \ln \det(\boldsymbol{\Sigma}_j^{-1}) - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_j)^\top \boldsymbol{\Sigma}_j^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_j) \right),$$

which can be easily maximized w.r.t. $\boldsymbol{\theta}$.

$$\hat{\pi}_j := \frac{1}{n} \sum_{i=1}^n w_{i,j}$$
$$\hat{\boldsymbol{\mu}}_j := \frac{1}{n\hat{\pi}_j} \sum_{i=1}^n w_{i,j} \boldsymbol{x}_i$$
$$\widehat{\boldsymbol{\Sigma}}_j := \frac{1}{n\hat{\pi}_j} \sum_{i=1}^n w_{i,j} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_j)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_j)^\top.$$

Log-likelihood vs. complete log-likelihood

Three different functions of $\boldsymbol{\theta}$:

1. **Log-likelihood function**, given observed data:

$$\mathcal{L}(\boldsymbol{\theta}) = \ln P_{\boldsymbol{\theta}}(\text{observed data}).$$

2. **Complete log-likelihood function**, given observed **and unobserved** data:

$$\mathcal{L}_c(\boldsymbol{\theta}) = \ln P_{\boldsymbol{\theta}}(\text{observed data}, \text{unobserved data}).$$

⋆ Suppose we have some initial guess of parameters $\hat{\boldsymbol{\theta}}$.
Treat **unobserved data** as random variables.
Conditional distribution of unobserved data given observed data $P_{\hat{\boldsymbol{\theta}}}$:

$$P_{\hat{\boldsymbol{\theta}}}(\text{unobserved data} \mid \text{observed data}).$$

(E.g., distribution of "soft labels" given the observed data.)

3. **Expected complete log-likelihood function** given observed data:

$$\mathbb{E}_{\hat{\boldsymbol{\theta}}}\left[\mathcal{L}_c(\boldsymbol{\theta}) \mid \text{observed data}\right] = \mathbb{E}_{\hat{\boldsymbol{\theta}}}\left[\ln P_{\boldsymbol{\theta}}(\text{observed \& unobserved data}) \mid \text{observed data}\right]$$

(Expectation $\mathbb{E}_{\hat{\boldsymbol{\theta}}}[\cdots]$ is with respect to $P_{\hat{\boldsymbol{\theta}}}$ given observed data.)

Expected complete log-likelihood

Given initial guess of parameters $\hat{\boldsymbol{\theta}}$,

$$w_{i,j} := \mathbb{E}_{\hat{\boldsymbol{\theta}}}\left[\Phi_{i,j} \mid \boldsymbol{X}_i = \boldsymbol{x}_i \forall i \in [n]\right]$$

can be interpreted as predicted "soft labels".

Using these $w_{i,j}$ to form expected complete log-likelihood function:

$$\mathbb{E}_{\hat{\boldsymbol{\theta}}}\left[\sum_{i=1}^n \sum_{j=1}^K \Phi_{i,j} \left(\ln \pi_j + \frac{1}{2} \ln \det(\boldsymbol{\Sigma}_j^{-1}) - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_j)^\top \boldsymbol{\Sigma}_j^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_j) \right) \middle| \boldsymbol{X}_i = \boldsymbol{x}_i \forall i \in [n] \right]$$
$$= \sum_{i=1}^n \sum_{j=1}^K w_{i,j} \left(\ln \pi_j + \frac{1}{2} \ln \det(\boldsymbol{\Sigma}_j^{-1}) - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_j)^\top \boldsymbol{\Sigma}_j^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_j) \right).$$

This function of θ is easy to maximize!

(But why do we care about the expected complete log-likelihood?)

Expectation-Maximization (E-M)

E-M algorithm for Gaussian mixtures
Initialize $\hat{\boldsymbol{\theta}} = (\hat{\pi}_1, \hat{\boldsymbol{\mu}}_1, \widehat{\boldsymbol{\Sigma}}_1, \dots, \hat{\pi}_K, \hat{\boldsymbol{\mu}}_K, \widehat{\boldsymbol{\Sigma}}_K)$ somehow. Then repeat:

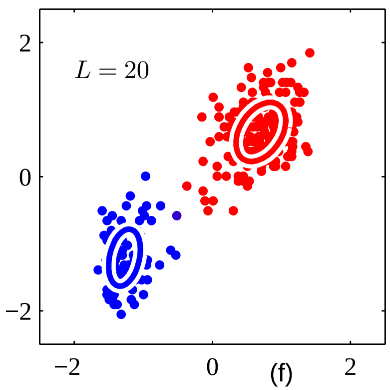
1. **E step**: expectation of "hidden variables" w.r.t. $P_{\hat{\boldsymbol{\theta}}}$ conditioned on data.
For each $i \in \{1, 2, \dots, n\}$ and $j \in \{1, 2, \dots, K\}$,

$$w_{i,j} := \frac{\hat{\pi}_j \cdot \sqrt{\det(\widehat{\boldsymbol{\Sigma}}_j^{-1})} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^\top \widehat{\boldsymbol{\Sigma}}_j^{-1} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j) \right)}{\sum_{j'=1}^K \hat{\pi}_{j'} \cdot \sqrt{\det(\widehat{\boldsymbol{\Sigma}}_{j'}^{-1})} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_{j'})^\top \widehat{\boldsymbol{\Sigma}}_{j'}^{-1} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_{j'}) \right)}$$

2. **M step**: maximize "expected complete log-likelihood" w.r.t. parameters.
For each $j \in \{1, 2, \dots, K\}$,

$$\hat{\pi}_j := \frac{1}{n} \sum_{i=1}^n w_{i,j}$$
$$\hat{\boldsymbol{\mu}}_j := \frac{1}{n\hat{\pi}_j} \sum_{i=1}^n w_{i,j} \boldsymbol{x}_i$$
$$\widehat{\boldsymbol{\Sigma}}_j := \frac{1}{n\hat{\pi}_j} \sum_{i=1}^n w_{i,j} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_j)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_j)^\top.$$

Sample run of the E-M algorithm



After **20 rounds** of E-M.

Using the E-M algorithm

E-M for Gaussian mixtures

1. **E step**: For each $i \in [n]$, $j \in [K]$,

$$w_{i,j} \;\propto\; \hat{\pi}_j \cdot p_{\hat{\boldsymbol{\mu}}_j, \widehat{\boldsymbol{\Sigma}}_j}(\boldsymbol{x}_i)$$

where $p_{\boldsymbol{\mu},\boldsymbol{\Sigma}}$ is the $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ pdf.

2. **M step**: For each $j \in [K]$,

$$\hat{\pi}_j \;:=\; \frac{1}{n}\sum_{i=1}^{n} w_{i,j}$$

$$\hat{\boldsymbol{\mu}}_j \;:=\; \frac{1}{n\hat{\pi}_j}\sum_{i=1}^{n} w_{i,j}\boldsymbol{x}_i$$

$$\widehat{\boldsymbol{\Sigma}}_j \;:=\; \frac{1}{n\hat{\pi}_j}\sum_{i=1}^{n} w_{i,j}(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_j)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_j)^{\top}.$$

Some details

- **No step sizes to tune!**
- **Initialization**: a bit of an art; both D^2-sampling and Lloyd's algorithm are reasonable.
- **Starved clusters**: problems can occur if $\hat{\pi}_j$ becomes too small (e.g., $\widehat{\boldsymbol{\Sigma}}_j$ could be near singular).

  Remove/replace such components.
- Log-likelihood of E-M iterates is **non-decreasing**; converges to a **stationary point**.

  ∴ Run E-M from many random initializations; pick the result with highest likelihood.

Derivation of E-M

E-M algorithm (Dempster, Laird, and Rubin, 1977)

E-M is a general algorithmic template for climbing log-likelihood objectives of models with **latent variables** (e.g., cluster assignments).

- What is the role of the expected complete log-likelihood?
- Why do parameters produced E-M iterations

$$\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(t)}$$

have increasing (or at least non-decreasing) log-likelihood

$$\mathcal{L}(\boldsymbol{\theta}^{(1)}) \;\le\; \mathcal{L}(\boldsymbol{\theta}^{(2)}) \;\le\; \cdots \;\le\; \mathcal{L}(\boldsymbol{\theta}^{(t)})\,?$$

Setting

Statistical model $\mathcal{P} = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$, where each $P_{\boldsymbol{\theta}} \in \mathcal{P}$ specifies distribution of observed variables $\boldsymbol{X}$ and latent variables $\boldsymbol{Y}$.

(For simplicity, assume they are discrete random variables.)

Here, if we want to consider models for iid samples, then we shall take $\mathcal{P}$ to be an appropriate "lifted" model (in the n-fold product form $\mathcal{P} = \mathcal{P}_0^n$ for a base model $\mathcal{P}_0$).

- So, if $\mathcal{P}$ is a "lifted" model, then $\boldsymbol{X}$ and $\boldsymbol{Y}$ are observed and latent variables for *all* n data points in the sample.

Expectation-Maximization

Initialize parameters $\boldsymbol{\theta}^{(1)} \in \Theta$ somehow.
For $t = 1, 2, \ldots$:

- **E step**: Construct expected complete log-likelihood function

$$\boldsymbol{\theta} \;\mapsto\; \mathbb{E}_{\boldsymbol{\theta}^{(t)}}[\mathcal{L}_c(\boldsymbol{\theta}) \mid \boldsymbol{X} = \boldsymbol{x}]$$

  where expectation (of latent variables) is with respect to distribution $P_{\boldsymbol{\theta}^{(t)}}$ given observed data $\boldsymbol{X} = \boldsymbol{x}$.

- **M step**: Choose $\boldsymbol{\theta}^{(t+1)}$ to maximize that function,

$$\boldsymbol{\theta}^{(t+1)} \;:=\; \arg\max_{\boldsymbol{\theta} \in \Theta} \; \mathbb{E}_{\boldsymbol{\theta}^{(t)}}[\mathcal{L}_c(\boldsymbol{\theta}) \mid \boldsymbol{X} = \boldsymbol{x}] \,.$$

Log-likelihood and lower-bounds

- Let $\boldsymbol{x}$ be the (observed) data.
- Let $q^{(t)}$ denote distribution of $\boldsymbol{Y}$ given $\boldsymbol{X} = \boldsymbol{x}$ when $(\boldsymbol{X}, \boldsymbol{Y}) \sim P_{\boldsymbol{\theta}^{(t)}}$:

$$q^{(t)}(\boldsymbol{y}) \;:=\; \frac{P_{\boldsymbol{\theta}^{(t)}}(\boldsymbol{x}, \boldsymbol{y})}{\sum_{\boldsymbol{y}' \in \mathcal{Y}} P_{\boldsymbol{\theta}^{(t)}}(\boldsymbol{x}, \boldsymbol{y}')} \,.$$

- Log-likelihood function $\mathcal{L}$:

$$\mathcal{L}(\boldsymbol{\theta}) \;=\; \ln\left(\sum_{\boldsymbol{y} \in \mathcal{Y}} q^{(t)}(\boldsymbol{y}) \cdot \frac{P_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y})}{q^{(t)}(\boldsymbol{y})}\right) \;=\; \ln\left(\mathbb{E}_{\boldsymbol{Y} \sim q^{(t)}}\left[\frac{P_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{Y})}{q^{(t)}(\boldsymbol{Y})}\right]\right) .$$

- Define **log-likelihood lower-bound** function $\mathcal{L}_{\mathrm{LB}}^{(t)}$ by

$$\mathcal{L}_{\mathrm{LB}}^{(t)}(\boldsymbol{\theta}) \;:=\; \mathbb{E}_{\boldsymbol{Y} \sim q^{(t)}}\left[\ln\left(\frac{P_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{Y})}{q^{(t)}(\boldsymbol{Y})}\right)\right] .$$

  (Here, $\mathbb{E}_{\boldsymbol{Y} \sim q^{(t)}}[\cdots]$ means expectation conditioned on $\boldsymbol{X} = \boldsymbol{x}$ under $q^{(t)}$.)

Relation to expected complete log-likelihood

Log-likelihood lower-bound function:

$$\begin{aligned}
\mathcal{L}_{\mathrm{LB}}^{(t)}(\boldsymbol{\theta}) \;&:=\; \mathbb{E}_{\boldsymbol{Y} \sim q^{(t)}}\left[\ln\left(\frac{P_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{Y})}{q^{(t)}(\boldsymbol{Y})}\right)\right] \\
&=\; \underbrace{\mathbb{E}_{\boldsymbol{Y} \sim q^{(t)}}[\ln P_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{Y})]}_{\substack{\text{Expected complete log-likelihood} \\ \mathbb{E}_{\boldsymbol{\theta}^{(t)}}[\mathcal{L}_c(\boldsymbol{\theta}) \mid \text{observed data}]}} \;+\; (\text{stuff not depending on } \boldsymbol{\theta}) \,.
\end{aligned}$$

Therefore, maximizing $\mathcal{L}_{\mathrm{LB}}^{(t)}$ is the same as maximizing expected complete log-likelihood

$$\boldsymbol{\theta} \;\mapsto\; \mathbb{E}_{\boldsymbol{\theta}^{(t)}}\left[\mathcal{L}_c(\boldsymbol{\theta}) \mid \text{observed data}\right] .$$

Lower-bound property

Jensen's inequality: for any *concave* function g and random variable Z,

$$g(\mathbb{E}[Z]) \;\geq\; \mathbb{E}[g(Z)]\,.$$

Since natural logarithm is concave,

$$\mathcal{L}(\boldsymbol{\theta}) \;=\; \ln\left(\mathbb{E}_{\boldsymbol{Y}\sim q^{(t)}}\left[\frac{P_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{Y})}{q^{(t)}(\boldsymbol{Y})}\right]\right) \;\geq\; \mathbb{E}_{\boldsymbol{Y}\sim q^{(t)}}\left[\ln\left(\frac{P_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{Y})}{q^{(t)}(\boldsymbol{Y})}\right)\right] \;=\; \mathcal{L}_{\mathrm{LB}}^{(t)}(\boldsymbol{\theta})\,.$$

Moreover, we have $\mathcal{L}(\boldsymbol{\theta}^{(t)}) = \mathcal{L}_{\mathrm{LB}}^{(t)}(\boldsymbol{\theta}^{(t)})$:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{Y}\sim q^{(t)}}\left[\ln\left(\frac{P_{\boldsymbol{\theta}^{(t)}}(\boldsymbol{x},\boldsymbol{Y})}{q^{(t)}(\boldsymbol{Y})}\right)\right] \;&=\; \mathbb{E}_{\boldsymbol{Y}\sim q^{(t)}}\left[\ln\left(\sum_{\boldsymbol{y}\in\mathcal{Y}} P_{\boldsymbol{\theta}^{(t)}}(\boldsymbol{x},\boldsymbol{y})\right)\right] \\
&=\; \ln\left(\sum_{\boldsymbol{y}\in\mathcal{Y}} P_{\boldsymbol{\theta}^{(t)}}(\boldsymbol{x},\boldsymbol{y})\right) \;=\; \mathcal{L}(\boldsymbol{\theta}^{(t)})\,.
\end{aligned}$$

E-M in terms of log-likelihood lower-bound

In t-th iteration of E-M:

- **E step**: Construct log-likelihood lower-bound function $\mathcal{L}_{\mathrm{LB}}^{(t)}$ via distribution $q^{(t)}$,

$$q^{(t)}(\boldsymbol{y}) \;:=\; \frac{P_{\boldsymbol{\theta}^{(t)}}(\boldsymbol{x},\boldsymbol{y})}{\sum_{\boldsymbol{y}'\in\mathcal{Y}} P_{\boldsymbol{\theta}^{(t)}}(\boldsymbol{x},\boldsymbol{y}')}\,,$$

  so that lower-bound is tight at $\boldsymbol{\theta}^{(t)}$.

- **M step**: Choose $\boldsymbol{\theta}^{(t+1)}$ to maximize $\mathcal{L}_{\mathrm{LB}}^{(t)}$:

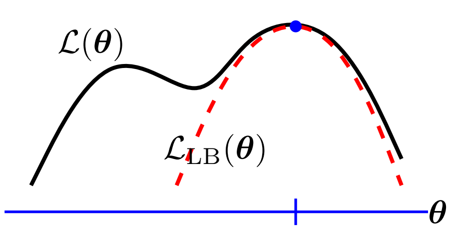
$$\boldsymbol{\theta}^{(t+1)} \;:=\; \arg\max_{\boldsymbol{\theta}\in\Theta} \mathcal{L}_{\mathrm{LB}}^{(t)}(\boldsymbol{\theta})\,.$$

Theorem

In t-th iteration of E-M, we have

$$\mathcal{L}(\boldsymbol{\theta}^{(t)}) \;=\; \mathcal{L}_{\mathrm{LB}}^{(t)}(\boldsymbol{\theta}^{(t)}) \;\leq\; \mathcal{L}_{\mathrm{LB}}^{(t)}(\boldsymbol{\theta}^{(t+1)}) \;\leq\; \mathcal{L}(\boldsymbol{\theta}^{(t+1)})\,.$$

Constructing and maximizing $\mathcal{L}_{\mathrm{LB}}$



M step: choose $\hat{\boldsymbol{\theta}}$ to maximize $\mathcal{L}_{\mathrm{LB}}$.

Key takeaways

1. **Mixture models**: similar to generative models for classification, except class labels are not observed.
   - Important example: **Gaussian mixture models**.
2. **E-M algorithm for Gaussian mixture models.**
3. **Recipe to derive E-M algorithm** for a general latent variable model:
   - **E step**: use conditional distribution of latent variables (given observed variables) under $P_{\boldsymbol{\theta}^{(t)}}$ to form "expected complete log-likelihood" function.
   - **M step**: maximize the "expected complete log-likelihood" function.
   - (We'll see more examples next time.)
4. **Key properties of E-M**:
   - "Expected complete log-likelihood" function is **lower-bound on log-likelihood function** that is tight at $\boldsymbol{\theta}^{(t)}$.
   - Log-likelihoods of E-M iterates are **non-decreasing**.

More latent variable models

Latent variable models

- Statistical model $\mathcal{P} = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ with both *observed variables* $\boldsymbol{X}$ and *latent variables* $\boldsymbol{Z}$.
- Would like to use MLE for parameter estimation, but this is often computationally difficult.
- Typically resort to using E-M to find parameters with high likelihood (though maybe not the highest possible).

(Actually, there are many options besides MLE . . .)

Log-likelihood vs. complete log-likelihood

1. **Log-likelihood function**, given observed data $\boldsymbol{x}$:
$$\mathcal{L}(\boldsymbol{\theta}) \;=\; \ln\left(\sum_{\boldsymbol{z}\in\mathcal{Z}} P_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{z})\right).$$
(If $\boldsymbol{Z}$ is continuous r.v., replace sum with integral . . .)
2. **Complete log-likelihood function**, given observed data $\boldsymbol{x}$ and unobserved data $\boldsymbol{z}$:
$$\mathcal{L}_c(\boldsymbol{\theta}) \;=\; \ln P_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{z}).$$
3. **Expected complete log-likelihood function** given observed data $\boldsymbol{x}$, treating unobserved data as random $\boldsymbol{Z}$:
$$\mathbb{E}_{\hat{\boldsymbol{\theta}}}\left[\mathcal{L}_c(\boldsymbol{\theta})\right] \;=\; \mathbb{E}_{\hat{\boldsymbol{\theta}}}\left[\ln P_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{Z}) \mid \boldsymbol{X}=\boldsymbol{x}\right].$$
(Expectation $\mathbb{E}_{\hat{\boldsymbol{\theta}}}[\cdots]$ is with respect to $P_{\hat{\boldsymbol{\theta}}}$ given $\boldsymbol{X}=\boldsymbol{x}$.)

Expectation-Maximization

Initialize parameters $\boldsymbol{\theta}^{(1)} \in \Theta$ somehow.
For $t = 1, 2, \ldots$:

- **E step**: Construct <u>expected</u> complete log-likelihood function
$$\boldsymbol{\theta} \;\mapsto\; \mathbb{E}_{\boldsymbol{\theta}^{(t)}}[\mathcal{L}_c(\boldsymbol{\theta}) \mid \boldsymbol{X}=\boldsymbol{x}]$$
where expectation (of latent variables) is with respect to distribution $P_{\boldsymbol{\theta}^{(t)}}$ given observed data $\boldsymbol{X}=\boldsymbol{x}$.
- **M step**: Choose $\boldsymbol{\theta}^{(t+1)}$ to <u>maximize</u> that function,
$$\boldsymbol{\theta}^{(t+1)} \;:=\; \operatorname*{arg\,max}_{\boldsymbol{\theta}\in\Theta}\; \mathbb{E}_{\boldsymbol{\theta}^{(t)}}[\mathcal{L}_c(\boldsymbol{\theta}) \mid \boldsymbol{X}=\boldsymbol{x}].$$

Example instantiations of E-M

1. Mixture models (last time)
2. Mechanical Turk model
3. Conditional mixture models
4. ...

Mechanical Turk model

Amazon Mechanical Turk

Passonneau and Carpenter (TACL 2014): asked Mechanical Turk workers (presumably humans) to label words by their "word sense".

- **Items** ("human intelligence tasks"): English words with multiple possible meanings, as they appear in natural sentences.
- **Labels**: the correct meanings in the given contexts.
- Some workers are **adept** at this task (likely to give correct answer), some are **inept** (as good as random guessing), others are **malicious** (likely to give wrong answer).
- For binary labels, can be difficult to distinguish adept from malicious, but can at least hope to ignore inept workers.

  If assume more adept workers than malicious workers, can use (weighted) majority vote over non-inept workers' labels.
- Modeling worker accuracies → labeled data set with label "reliabilities".

Mechanical Turk model (Dawid and Skene, 1979)

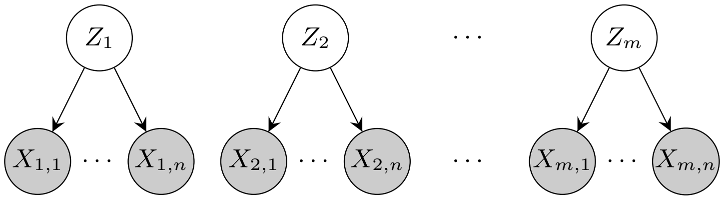
- **Observed**: predicted labels on m items from n workers $\{x_{i,j}\}_{i\in[m],j\in[n]}$.
- **Hidden**: correct labels $\{z_i\}_{i=1}^m$ for all m items.
- **Model**: $\mathcal{P} = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$, where $\boldsymbol{\theta} = (\pi_1, \pi_2, \ldots, \pi_m, p_1, p_2, \ldots, p_n)$.
  - Binary labels ($\{0,1\}$).
  - Data for items $\{(Z_i, X_{i,1}, X_{i,2}, \ldots, X_{i,n})\}_{i=1}^m$ are independent.
  - Nature determines correct label for item i:

    $$P_{\boldsymbol{\theta}}(Z_i = 1) \;=\; 1 - P_{\boldsymbol{\theta}}(Z_i = 0) \;=\; \pi_i \,.$$
  - Conditioned on Z_i, predicted labels $\{X_{i,j}\}_{j=1}^n$ from workers on item i are independent.
  - Worker j is correct with probability p_j:

    $$P_{\boldsymbol{\theta}}(X_{i,j} = z \mid Z_i = z) \;=\; p_j \,.$$
- **Notation**:
  - $\boldsymbol{Z} = (Z_1, \ldots, Z_m)$ (all correct labels);
  - $\boldsymbol{X} = (X_{1,1}, \ldots, X_{m,n})$ (all predicted labels);
  - $\boldsymbol{X}_i = (X_{i,1}, \ldots, X_{i,n})$ (predicted labels for item i).

Aside: graphical models

(Directed) graphical model: use (directed) graph structure over random variables to represent conditional independence assumptions.



Semantics

- Observed variables shaded, latent variables not shaded.
- **Disjoint connected components** are **independent** of each other.
  E.g., $(Z_1, \boldsymbol{X}_1) \perp\!\!\!\perp (Z_2, \boldsymbol{X}_2) \perp\!\!\!\perp \cdots \perp\!\!\!\perp (Z_m, \boldsymbol{X}_m)$.
- If connected component is **directed tree** (i.e., every node has ≤ 1 parent), **conditioning on r.v.** is akin to **removing it from the graph**.
  E.g., $X_{i,1} \perp\!\!\!\perp X_{i,2} \perp\!\!\!\perp \cdots \perp\!\!\!\perp X_{i,n} \mid Z_i = z$.
- Other rules more involved (e.g., for vertices with multiple parents).

Complete log-likelihood

Complete log-likelihood of $\boldsymbol{\theta} = (\pi_1, \pi_2, \ldots, \pi_m, p_1, p_2, \ldots, p_n)$ given unobserved data $\boldsymbol{z}$ and observed data $\boldsymbol{x}$:

(Use *conditional independence* properties to *factor* the likelihood.)

$$\begin{aligned}
\mathcal{L}_c(\boldsymbol{\theta}) &= \ln \prod_{i=1}^m P_{\boldsymbol{\theta}}(Z_i = z_i) \prod_{j=1}^n P_{\boldsymbol{\theta}}(X_{i,j} = x_{i,j} \mid Z_i = z_i) \\
&= \ln \prod_{i=1}^m \pi_i^{z_i}(1-\pi_i)^{1-z_i} \prod_{j=1}^n p_j^{(1-x_{i,j})(1-z_i)+x_{i,j}z_i}(1-p_j)^{(1-x_{i,j})z_i+x_{i,j}(1-z_i)} \\
&= \sum_{i=1}^m \big[z_i \ln \pi_i + (1-z_i)\ln(1-\pi_i)\big] \\
&\quad + \sum_{i=1}^m \sum_{j=1}^n \big[(1-x_{i,j})(1-z_i) + x_{i,j}z_i\big] \ln p_j \\
&\quad + \sum_{i=1}^m \sum_{j=1}^n \big[(1-x_{i,j})z_i + x_{i,j}(1-z_i)\big] \ln(1-p_j)
\end{aligned}$$

Expected complete log-likelihood

Assume we have some initial parameters $\hat{\boldsymbol{\theta}} = (\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_m, \hat{p}_1, \hat{p}_2, \ldots, \hat{p}_n)$.
Expected complete log-likelihood of $\boldsymbol{\theta}$ w.r.t. $P_{\hat{\boldsymbol{\theta}}}$ conditioned on $\boldsymbol{X} = \boldsymbol{x}$:

$$\begin{aligned}
\mathbb{E}_{\hat{\boldsymbol{\theta}}}\big[\mathcal{L}_c(\boldsymbol{\theta}) \mid \boldsymbol{X} = \boldsymbol{x}\big] &= \sum_{i=1}^m \big[w_i \ln \pi_i + (1-w_i)\ln(1-\pi_i)\big] \\
&\quad + \sum_{i=1}^m \sum_{j=1}^n \big[(1-x_{i,j})(1-w_i) + x_{i,j}w_i\big] \ln p_j \\
&\quad + \sum_{i=1}^m \sum_{j=1}^n \big[(1-x_{i,j})w_i + x_{i,j}(1-w_i)\big] \ln(1-p_j)
\end{aligned}$$

where

$$\begin{aligned}
w_i &:= P_{\hat{\boldsymbol{\theta}}}\big(Z_i = 1 \mid \boldsymbol{X}_i = \boldsymbol{x}_i\big) = \frac{P_{\hat{\boldsymbol{\theta}}}(Z_i = 1, \boldsymbol{X}_i = \boldsymbol{x}_i)}{P_{\hat{\boldsymbol{\theta}}}(Z_i = 1, \boldsymbol{X}_i = \boldsymbol{x}_i) + P_{\hat{\boldsymbol{\theta}}}(Z_i = 0, \boldsymbol{X}_i = \boldsymbol{x}_i)} \\
&= \frac{\hat{\pi}_i \prod_{j=1}^n \hat{p}_j^{x_{i,j}}(1-\hat{p}_j)^{1-x_{i,j}}}{\hat{\pi}_i \prod_{j=1}^n \hat{p}_j^{x_{i,j}}(1-\hat{p}_j)^{1-x_{i,j}} + (1-\hat{\pi}_i)\prod_{j=1}^n \hat{p}_j^{1-x_{i,j}}(1-\hat{p}_j)^{x_{i,j}}}.
\end{aligned}$$

Interpreting w_i

$$\frac{w_i}{1-w_i} = \frac{P_{\hat{\boldsymbol{\theta}}}\big(Z_i = 1 \mid \boldsymbol{X}_i = \boldsymbol{x}_i\big)}{P_{\hat{\boldsymbol{\theta}}}\big(Z_i = 0 \mid \boldsymbol{X}_i = \boldsymbol{x}_i\big)} = \frac{\hat{\pi}_i}{1-\hat{\pi}_i} \prod_{j=1}^n \left(\frac{\hat{p}_j}{1-\hat{p}_j}\right)^{2x_{i,j}-1}.$$

- Start with $\frac{\hat{\pi}_i}{1-\hat{\pi}_i}$: "prior" odds ratio.
- Perfectly inept workers $\hat{p}_j = 1/2$: **contributes nothing**.
- Adept workers $\hat{p}_j > 1/2$:
  - If $x_{i,j} = 1$: **increase ratio**.
  - If $x_{i,j} = 0$: **decrease ratio**.
- Malicious workers $\hat{p}_j < 1/2$:
  - If $x_{i,j} = 1$: **decrease ratio**.
  - If $x_{i,j} = 0$: **increase ratio**.

Maximizing $\mathbb{E}_{\hat{\boldsymbol{\theta}}}[\mathcal{L}_c(\boldsymbol{\theta}) \mid \boldsymbol{X} = \boldsymbol{x}]$

Function $\boldsymbol{\theta} \mapsto \mathbb{E}_{\hat{\boldsymbol{\theta}}}[\mathcal{L}_c(\boldsymbol{\theta}) \mid \boldsymbol{X} = \boldsymbol{x}]$ is concave w.r.t. $\boldsymbol{\theta}$;

$$\begin{aligned} \frac{\partial}{\partial \pi_i} \mathbb{E}_{\hat{\boldsymbol{\theta}}}\left[\mathcal{L}_c(\boldsymbol{\theta}) \mid \boldsymbol{X} = \boldsymbol{x}\right] &= \frac{w_i}{\pi_i} - \frac{1 - w_i}{1 - \pi_i}, \\ \frac{\partial}{\partial p_j} \mathbb{E}_{\hat{\boldsymbol{\theta}}}\left[\mathcal{L}_c(\boldsymbol{\theta}) \mid \boldsymbol{X} = \boldsymbol{x}\right] &= \frac{\sum_{i=1}^m (1 - x_{i,j})(1 - w_i) + x_{i,j} w_i}{p_j} \\ &\quad - \frac{\sum_{i=1}^m (1 - x_{i,j}) w_i + x_{i,j}(1 - w_i)}{1 - p_j}. \end{aligned}$$

Partial derivatives are zero when

$$\begin{aligned} \pi_i &= w_i, \\ p_j &= \frac{1}{m} \sum_{i=1}^m \left\{ w_i x_{i,j} + (1 - w_i)(1 - x_{i,j}) \right\}. \end{aligned}$$

Interpreting maximizers

Maximizers of expected complete log-likelihood function:

$$\begin{aligned} \pi_i &= w_i, \\ p_j &= \frac{1}{m} \sum_{i=1}^m \left\{ w_i x_{i,j} + (1 - w_i)(1 - x_{i,j}) \right\}. \end{aligned}$$

- $\pi_i = w_i$: of course.
- Now pretend w_i are true "soft" labels of items.
- p_j: "soft" accuracy of worker j based on predicted labels across all items.

E-M algorithm

Initialize $\hat{\boldsymbol{\theta}} = (\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_m, \hat{p}_1, \hat{p}_2, \ldots, \hat{p}_n)$.
Repeat:

- **E step**:

$$w_i := \frac{\hat{\pi}_i \prod_{j=1}^n \hat{p}_j^{x_{i,j}} (1 - \hat{p}_j)^{1 - x_{i,j}}}{\hat{\pi}_i \prod_{j=1}^n \hat{p}_j^{x_{i,j}} (1 - \hat{p}_j)^{1 - x_{i,j}} + (1 - \hat{\pi}_i) \prod_{j=1}^n \hat{p}_j^{1 - x_{i,j}} (1 - \hat{p}_j)^{x_{i,j}}}$$

  for all $i = 1, 2, \ldots, m$.

- **M step**:

$$\begin{aligned} \hat{\pi}_i &:= w_i \quad \text{for all } i = 1, 2, \ldots, m, \\ \hat{p}_j &:= \frac{1}{m} \sum_{i=1}^m \left\{ w_i x_{i,j} + (1 - w_i)(1 - x_{i,j}) \right\} \quad \text{for all } j = 1, 2, \ldots, n. \end{aligned}$$

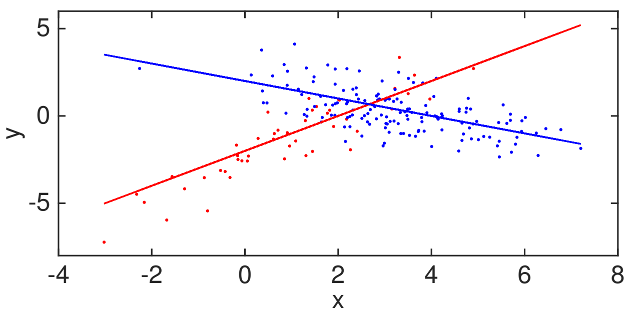
Can extend to handle multi-class labels, label-dependent worker abilities, etc.

Conditional mixture model

Conditional mixture model

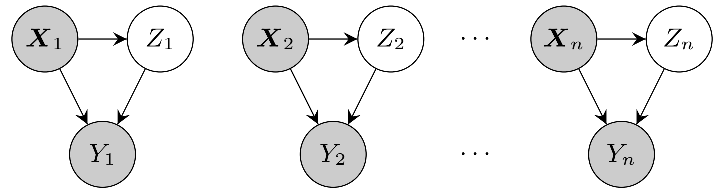
Value of $\boldsymbol{x}$ (probabilistically) determines which regression model y follows: e.g.,

$$\begin{aligned} y &= \langle \boldsymbol{\beta}_0, \boldsymbol{x} \rangle + \text{noise}\,, \\ \text{or} \quad y &= \langle \boldsymbol{\beta}_1, \boldsymbol{x} \rangle + \text{noise}\,. \end{aligned}$$



Mixture of two linear regressions

- **Observed**: labeled data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ from $\mathbb{R}^d \times \mathbb{R}$.
- **Hidden**: hidden bits $\{z_i\}_{i=1}^n$ from $\{0, 1\}$.
- **Model**: $\mathcal{P} = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$, where $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d$.
  - $\{(\boldsymbol{X}_i, Y_i, Z_i)\}_{i=1}^n$ are iid.
  - Hidden bit for i-th data point depends on $\boldsymbol{X}_i$:
    $$Z_i \mid \boldsymbol{X}_i = \boldsymbol{x}_i \;\sim\; \mathrm{Bern}\big(\mathrm{logistic}(\langle \boldsymbol{\alpha}, \boldsymbol{x}_i \rangle)\big)$$
    where $\mathrm{logistic}(t) = 1/(1 + e^{-t})$. (Note: $1 - \mathrm{logistic}(t) = \mathrm{logistic}(-t)$.)
  - Hidden bit determines which regression coefficients to use:
    $$Y_i \mid \boldsymbol{X}_i = \boldsymbol{x}_i,\, Z_i = z_i \;\sim\; \mathrm{N}\big(\langle (1 - z_i)\boldsymbol{\beta}_0 + z_i \boldsymbol{\beta}_1, \boldsymbol{x}_i \rangle, 1\big)\,.$$



Complete log-(conditional) likelihood

$$\begin{aligned} \mathcal{L}_c(\boldsymbol{\theta}) &= \ln \prod_{i=1}^n P_{\boldsymbol{\theta}}(Z_i = z_i \mid \boldsymbol{X}_i = \boldsymbol{x}_i) \cdot p_{\boldsymbol{\theta}}(y_i \mid \boldsymbol{X}_i = \boldsymbol{x}_i,\, Z_i = z_i) \\ &= \sum_{i=1}^n \Big(z_i \ln \mathrm{logistic}(\langle \boldsymbol{\alpha}, \boldsymbol{x}_i \rangle) + (1 - z_i) \ln \mathrm{logistic}(-\langle \boldsymbol{\alpha}, \boldsymbol{x}_i \rangle) \\ &\qquad - \frac{1}{2}\big(y_i - \langle (1 - z_i)\boldsymbol{\beta}_0 + z_i \boldsymbol{\beta}_1, \boldsymbol{x}_i \rangle\big)^2 \Big) + (\text{stuff not involving } \boldsymbol{\theta})\,. \end{aligned}$$

Given $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\beta}}_1)$,

$$\begin{aligned} w_i &:= \mathbb{E}_{\hat{\boldsymbol{\theta}}}\big[Z_i \mid \boldsymbol{X}_i = \boldsymbol{x}_i,\, Y_i = y_i\big] = P_{\hat{\boldsymbol{\theta}}}\big(Z_i = 1 \mid \boldsymbol{X}_i = \boldsymbol{x}_i,\, Y_i = y_i\big) \\ &= \frac{\mathrm{logistic}(\langle \hat{\boldsymbol{\alpha}}, \boldsymbol{x}_i \rangle) \cdot e^{-\frac{1}{2}\left(y_i - \langle \hat{\boldsymbol{\beta}}_1, \boldsymbol{x}_i \rangle\right)^2}}{\mathrm{logistic}(\langle \hat{\boldsymbol{\alpha}}, \boldsymbol{x}_i \rangle) \cdot e^{-\frac{1}{2}\left(y_i - \langle \hat{\boldsymbol{\beta}}_1, \boldsymbol{x}_i \rangle\right)^2} + \mathrm{logistic}(-\langle \hat{\boldsymbol{\alpha}}, \boldsymbol{x}_i \rangle) \cdot e^{-\frac{1}{2}\left(y_i - \langle \hat{\boldsymbol{\beta}}_0, \boldsymbol{x}_i \rangle\right)^2}}\,. \end{aligned}$$

Maximizing $\mathbb{E}_{\hat{\boldsymbol{\theta}}}[\mathcal{L}_c(\boldsymbol{\theta}) \mid \boldsymbol{X} = \boldsymbol{x},\, \boldsymbol{Y} = \boldsymbol{y}]$

$$\begin{aligned} \mathcal{L}_c(\boldsymbol{\theta}) &= \sum_{i=1}^n \Big(z_i \ln \mathrm{logistic}(\langle \boldsymbol{\alpha}, \boldsymbol{x}_i \rangle) + (1 - z_i) \ln \mathrm{logistic}(-\langle \boldsymbol{\alpha}, \boldsymbol{x}_i \rangle) \\ &\qquad - \frac{1}{2}\big(y_i - \langle (1 - z_i)\boldsymbol{\beta}_0 + z_i \boldsymbol{\beta}_1, \boldsymbol{x}_i \rangle\big)^2 \Big) + (\text{ignorable})\,. \end{aligned}$$

Replace z_i with r.v. Z_i, and take expectations:

$$\begin{aligned} &\mathbb{E}_{\hat{\boldsymbol{\theta}}}\big[\mathcal{L}_c(\boldsymbol{\theta}) \mid \boldsymbol{X} = \boldsymbol{x},\, \boldsymbol{Y} = \boldsymbol{y}\big] \\ &= \sum_{i=1}^n \Big(w_i \ln \mathrm{logistic}(\langle \boldsymbol{\alpha}, \boldsymbol{x}_i \rangle) + (1 - w_i) \ln \mathrm{logistic}(-\langle \boldsymbol{\alpha}, \boldsymbol{x}_i \rangle) \\ &\qquad - \frac{1 - w_i}{2}\big(y_i - \langle \boldsymbol{\beta}_0, \boldsymbol{x}_i \rangle\big)^2 - \frac{w_i}{2}\big(y_i - \langle \boldsymbol{\beta}_1, \boldsymbol{x}_i \rangle\big)^2 \Big) + (\text{ignorable})\,. \end{aligned}$$

- Maximizing w.r.t. $\boldsymbol{\alpha}$: weighted logistic regression.
- Maximizing w.r.t. $\boldsymbol{\beta}_0$: weighted linear regression.
- Maximizing w.r.t. $\boldsymbol{\beta}_1$: weighted linear regression.

E-M algorithm

Initialize $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\beta}}_1)$.
Repeat:

- **E step**:

$$w_i := \frac{\text{logistic}(\langle \hat{\boldsymbol{\alpha}}, \boldsymbol{x}_i \rangle) \cdot e^{-\frac{1}{2}\left(y_i - \langle \hat{\boldsymbol{\beta}}_1, \boldsymbol{x}_i \rangle\right)^2}}{\text{logistic}(\langle \hat{\boldsymbol{\alpha}}, \boldsymbol{x}_i \rangle) \cdot e^{-\frac{1}{2}\left(y_i - \langle \hat{\boldsymbol{\beta}}_1, \boldsymbol{x}_i \rangle\right)^2} + \text{logistic}(-\langle \hat{\boldsymbol{\alpha}}, \boldsymbol{x}_i \rangle) \cdot e^{-\frac{1}{2}\left(y_i - \langle \hat{\boldsymbol{\beta}}_0, \boldsymbol{x}_i \rangle\right)^2}}$$

  for all $i = 1, 2, \ldots, n$.

- **M step**: Solve a weighted logistic regression problem, and two weighted least squares problems to get new $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\beta}}_1)$.

Key takeaways

1. More examples of E-M algorithm for latent variable models.
   - Need to mind conditional independence assumptions. (Graphical models can help with this.)
2. Sometimes E- and M-steps are not available in closed-form.