Expectation Maxmization of Gaussian Mixture Models

James A. Perez

2019, April 25

Outline

- Overview:
  1. Commingling Analysis
  2. Gaussian Mixture Model (GMM)
  3. Expectation Maximization Algorithm
  4. Code Demonstration
  5. Results / Issues

Problem Statement

- Using simulated quantitative phenotype trait (QTP) measurements from N independent individuals $X = (x_1, x_2, ..., x_n)$, estimate the parameters of the k component denisties θ_k generative for each quantitative trait x_i.

Problem Statement

- Using simulated quantitative phenotype trait (QTP) measurements from N independent individuals $X = (x_1, x_2, ..., x_n)$, estimate the parameters of the k component denisties θ_k generative for each quantitative trait x_i.

Assumptions:

1. Mixture densities are members of a Gaussian location family (common variance).
2. Biallelic genotype forms a partiton in the sample space of X.

Gaussian Mixture Model

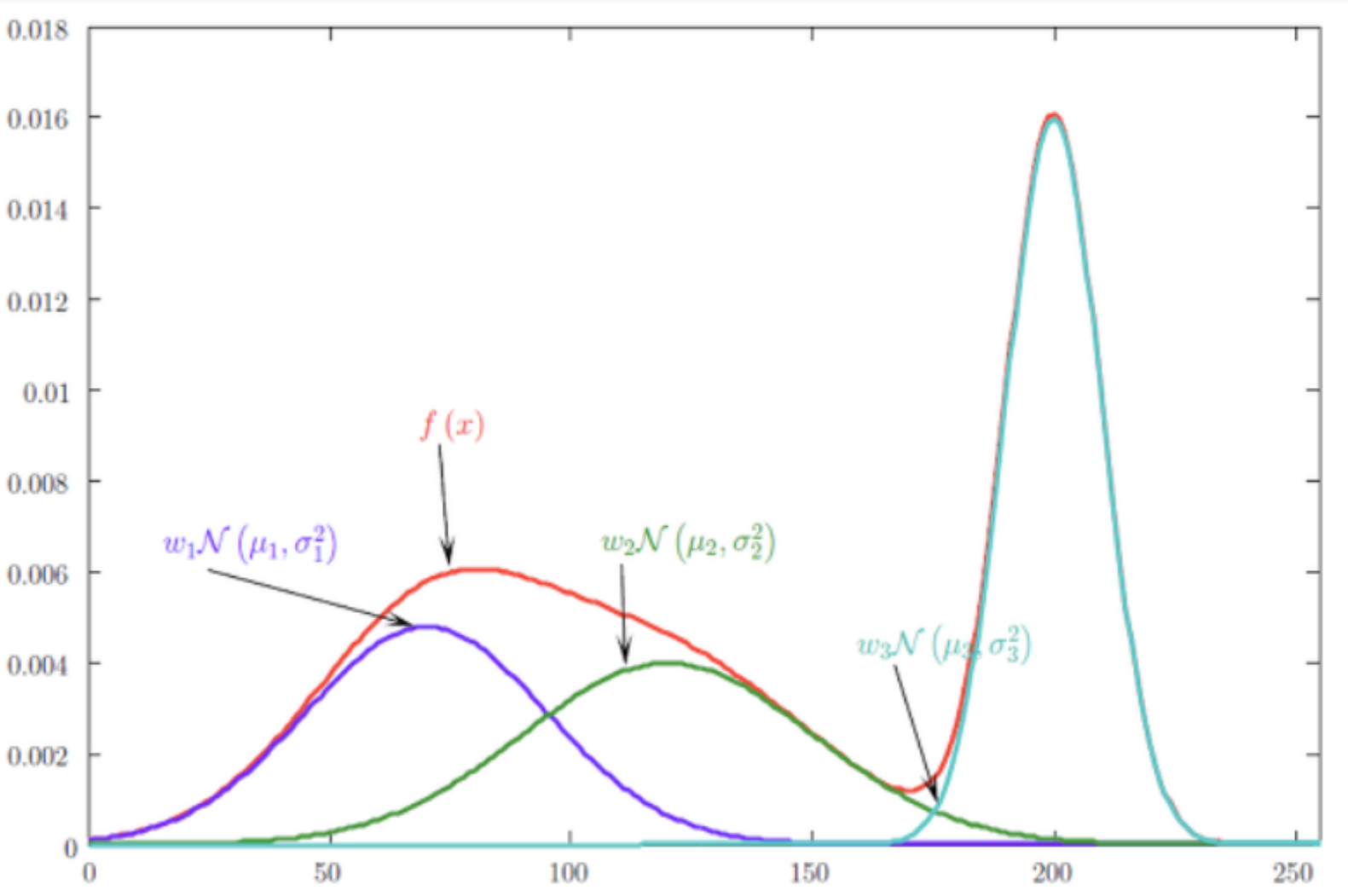

Figure 1:

Gaussian Mixture Model

- Also known as a disribution of "contaminated normals"

Given by,

$$f(x) = \alpha_1 f_1(x) + \alpha_2 f_2(x) + ... + \alpha_k f_k(x)$$

Gaussian Mixture Model

- Also known as a disribution of "contaminated normals"

Given by,

$$f(x) = \alpha_1 f_1(x) + \alpha_2 f_2(x) + ... + \alpha_k f_k(x)$$

- α_k = Probability that any realization $X = x_i$ was derived from density k
- We call this a mixture component probability.

Gaussian Mixture Model

$$f(x) = \alpha_1 f_1(x) + \alpha_2 f_2(x) + \alpha_k f_k(x)$$

- Expectation:

$$E[X] = \sum_{i=1}^{k} \alpha_i \int_{-\infty}^{\infty} x f_i(x) = \sum_{i=1}^{k} \alpha_i \mu_i = \bar{\mu}$$

- Variance:

$$Var(X) = \sum_{i=1}^{k} \alpha_i \sigma_i^2 + \sum_{i=1}^{k} \alpha_i (\mu_i - \bar{\mu})^2$$

Expectation Maximization (EM)

- The mixture density problem is one of the most widely used applications of the EM algorithm.

Expectation Maximization (EM)

- The mixture density problem is one of the most widely used applications of the EM algorithm.
- The probablility model we wish to maximize is,

$$p(x|\Theta) = \sum_{m=1}^{K} \alpha_m p_m(x|\theta_m)$$

- where $\Theta = (\alpha_1, ..., \alpha_k, \theta_1, ..., \theta_k)$ such that $\sum_{m=1}^{K} \alpha_m = 1$

Expectation Maximization (EM)

The incomplete log likelihood is given by,

$$log(L(\Theta|X)) = log \prod_{i=1}^{N} p(x_i|\Theta) = \sum_{i}^{N} log \left(\sum_{m=1}^{K} \alpha_m p_m(x_i|\theta_m) \right)$$

Expectation Maximization (EM)

The incomplete log likelihood is given by,

$$log(L(\Theta|X)) = log \prod_{i=1}^{N} p(x_i|\Theta) = \sum_{i}^{N} log \left(\sum_{m=1}^{K} \alpha_m p_m(x_i|\theta_m) \right)$$

- The log of a summation turns out to be very difficult to maximize using standard numerical techniques...

Expectation Maximization (EM)

The incomplete log likelihood is given by,

$$log(L(\Theta|X)) = log \prod_{i=1}^{N} p(x_i|\Theta) = \sum_{i}^{N} log \left(\sum_{m=1}^{K} \alpha_m p_m(x_i|\theta_m) \right)$$

- The log of a summation turns out to be very difficult to maximize using standard numerical techniques...
- How can we simplify the maximization computation?

Expectation Maximization (EM)

- Soln: We posit the existence of an unknown random indicator vector $Z_{ik} \in \{0, 1\}$ that informs us of which k^{th} component density was generative for each $X = x_i$

Expectation Maximization (EM)

- Soln: We posit the existence of an unknown random indicator vector $Z_{ik} \in \{0, 1\}$ that informs us of which k^{th} component density was generative for each $X = x_i$
- If we knew beforehand what the distribution of Z_k was, the likelihood would be:

Expectation Maximization (EM)

- Soln: We posit the existence of an unknown random indicator vector $Z_{ik} \in \{0, 1\}$ that informs us of which k^{th} component density was generative for each $X = x_i$
- If we knew beforehand what the distribution of Z_k was, the likelihood would be:

$$log(L(\Theta|X, Z_k) = \sum_{i=1}^{N} log(P(x_i|Z_{ik})P(Z_{ik} = 1)) = \sum_{i}^{N} log(\alpha_k p_k(x_i|\theta_k))$$

Expectation Maximization (EM)

- Soln: We posit the existence of an unknown random indicator vector $Z_{ik} \in \{0, 1\}$ that informs us of which k^{th} component density was generative for each $X = x_i$
- If we knew beforehand what the distribution of Z_k was, the likelihood would be:

$$log(L(\Theta|X, Z_k) = \sum_{i=1}^{N} log(P(x_i|Z_{ik})P(Z_{ik} = 1)) = \sum_{i}^{N} log(\alpha_k p_k(x_i|\theta_k))$$

- However, we of course don't know which k^{th}component was used to sample $X = x_i$

Expectation Maximization (EM)

- Soln: We posit the existence of an unknown random indicator vector $Z_{ik} \in \{0, 1\}$ that informs us of which k^{th} component density was generative for each $X = x_i$
- If we knew beforehand what the distribution of Z_k was, the likelihood would be:

$$log(L(\Theta|X, Z_k) = \sum_{i=1}^{N} log(P(x_i|Z_{ik})P(Z_{ik} = 1)) = \sum_{i}^{N} log(\alpha_k p_k(x_i|\theta_k))$$

- However, we of course don't know which k^{th}component was used to sample $X = x_i$
- Therefore, we must derive an expression for the distribution of the "unobserved" indicators $Z_k = (Z_{1k}, Z_{2k}, ..., Z_{nk})$.

Expectation Maximization (EM)

- The strategy is to pretend we know parameters for the mixture densities ("guesses") in order to derive the distribution of Z_k. to maximize our likelihood.

Expectation Maximization (EM)

- The strategy is to pretend we know parameters for the mixture densities ("guesses") in order to derive the distribution of Z_k. to maximize our likelihood.
- In that case, we pivot our interest to the following "complete" log likelihood:

$$\sum_{m=1}^{K} log(L(\Theta|X, Z_m))p(Z_m = 1|X, \Theta^g)$$

- where Θ^g is our first guess for the parameters.

Expectation Maximization (EM)

- Given Θ^g we can easily compute $p_k(x_i|\theta_k^g)$
- In addition, the mixture component probabilities α_k, can be thought of as Bayesian "priors".

Expectation Maximization (EM)

- Given Θ^g we can easily compute $p_k(x_i|\theta_k^g)$
- In addition, the mixture component probabilities α_k, can be thought of as Bayesian "priors".

Therefore, using Bayes's rule:

$$w_{ik} = p(Z_{ik} = 1|x_i, \Theta^g) = \frac{\alpha_k^g p_k(x_i|\theta_k^g)}{\sum_{m=1}^K \alpha_m^g p_m(x_i|\theta_m^g)}$$

where I denote w_{ik} as a membership weight of data point x_i in mixture denisty k.

Expectation Maximization (EM)

- Recall the complete log likelihood which I now denote $Q(\Theta|\Theta^g)$

$$Q(\Theta|\Theta^g) = \sum_{m=1}^{K} log(L(\Theta|X, Z_m))p(Z_m = 1|X, \Theta^g)$$

Expectation Maximization (EM)

- Recall the complete log likelihood which I now denote $Q(\Theta|\Theta^g)$

$$Q(\Theta|\Theta^g) = \sum_{m=1}^{K} log(L(\Theta|X, Z_m))p(Z_m = 1|X, \Theta^g)$$

- Now that we have the marginal denisty of Z_k from Bayes theorem, we can maximize the previous expression, denoted by

$$Q^*(\Theta|\Theta^g) = \underset{\Theta}{\text{argmax}} L(\Theta|X)$$

Expectation Maximization (EM)

- Recall the complete log likelihood which I now denote $Q(\Theta|\Theta^g)$

$$Q(\Theta|\Theta^g) = \sum_{m=1}^{K} log(L(\Theta|X, Z_m))p(Z_m = 1|X, \Theta^g)$$

- Now that we have the marginal denisty of Z_k from Bayes theorem, we can maximize the previous expression, denoted by

$$Q^*(\Theta|\Theta^g) = \underset{\Theta}{\text{argmax}} L(\Theta|X)$$

- The EM algorithm does this iteratively through i iterations

$$Q^{(i)} = \underset{\Theta}{\text{argmax}} Q(\Theta, \Theta^{(i-1)})$$

EM (E-Step)

- Each iteration has two steps:

EM (E-Step)

- Each iteration has two steps:

E-step: Compute w_{ik} for all data points $x_i 1 <= i <= N$ and mixture components $1 <= k <= K$ using the initial or old parameter vector $\Theta^{(i-1)}$. This yields the NxK matrix W, where each row sums to one.

EM (E-Step)

- Each iteration has two steps:

E-step: Compute w_{ik} for all data points $x_i 1 <= i <= N$ and mixture components $1 <= k <= K$ using the initial or old parameter vector $\Theta^{(i-1)}$. This yields the NxK matrix W, where each row sums to one.

$$w_{ij} = \frac{\alpha_k^{(i-1)} p_k \left(x_i | \theta_k^{(i-1)} \right)}{\sum_{m=1}^{K} \alpha_m^{(i-1)} p_m \left(x_i | \theta_m^{(i-1)} \right)} \quad (EQN1)$$

EM (M-Step)

M-step: Use the "priors" or membership weights w_{ij} to calculate the new parameter vector $\Theta^{(i)}$.

EM (M-Step)

M-step: Use the "priors" or membership weights w_{ij} to calculate the new parameter vector $\Theta^{(i)}$.

- Let $N_k = \sum_{i=1}^{N} w_{ik}$ be the effective number of data points assigned to component density k.

EM (M-Step)

M-step: Use the "priors" or membership weights w_{ij} to calculate the new parameter vector $\Theta^{(i)}$.

- Let $N_k = \sum_{i=1}^{N} w_{ik}$ be the effective number of data points assigned to component density k.

We have for the mixture components,

$$\alpha^{(i)} = \frac{N_k}{N} \quad (EQN2)$$

EM (M-Step)

M-step: Use the "priors" or membership weights w_{ij} to calculate the new parameter vector $\Theta^{(i)}$.

- Let $N_k = \sum_{i=1}^{N} w_{ik}$ be the effective number of data points assigned to component density k.

We have for the mixture components,

$$\alpha^{(i)} = \frac{N_k}{N} \quad (EQN2)$$

The updated means,

$$\mu^{(i)} = \frac{1}{N_k} \sum_{i=1}^{N} w_{ij} \cdot x_i, \quad 1 <= \text{k} <= \text{K}. \quad (EQN3)$$

EM (M-Step)

M-step: Use the "priors" or membership weights w_{ij} to calculate the new parameter vector $\Theta^{(i)}$.

- Let $N_k = \sum_{i=1}^{N} w_{ik}$ be the effective number of data points assigned to component density k.

We have for the mixture components,

$$\alpha^{(i)} = \frac{N_k}{N} \quad (EQN2)$$

The updated means,

$$\mu^{(i)} = \frac{1}{N_k} \sum_{i=1}^{N} w_{ij} \cdot x_i, \quad 1 <= \mathrm{k} <= \mathrm{K}. \quad (EQN3)$$

And *common variance*,

$$\sigma^{2(i)} = \frac{1}{N_k} \sum_{i=1}^{N} w_{ij} \cdot \left(x_i - \mu^{(i)}\right)^2 \quad 1 <= \mathrm{k} <= \mathrm{K}. \quad (EQN4)$$

Code demonstration

Results: Estimation Accuracy - tol=0.001

Set 1 (*itr=27*)					
	dq	μ_{ii}	μ_{ij}	μ_{jj}	σ^2
true	0.400	-0.600	0.000	0.800	0.100
est	0.412	0.586	0.007	0.765	0.009
D^2	1.52e-04	1.94e-04	4.32e-05	1.25e-03	8.25e-03
Set 2 (*itr=122*)					
true	0.100	0.200	0.300	0.700	0.500
est	0.575	0.178	0.314	0.314	0.235
D^2	2.26e-01	5.05e-04	1.86e-04	1.49e-01	7.03e-02
Set 3 (*itr=121*)					
true	0.050	0.200	0.300	0.700	0.500
est	0.576	0.248	0.285	0.285	0.242
D^2	2.77e-01	2.28e-03	2.30e-04	1.72e-01	6.63e-02

$\text{mse}_1 = 0.002$, $\text{mse}_2 = 0.089$, $\text{mse}_3 = 0.104$

Results: Estimation Accuracy - tol=0.00001

Set 3 (*itr=867*)					
true	0.050	0.200	0.300	0.700	0.500
est	0.577	0.231	0.293	0.293	0.241
D^2	2.78e-01	9.82e-04	4.63e-05	1.65e-01	6.70e-02

$\text{mse}_3 = 0.102$

Issues

1. EM has trouble converging for small disease allele frequencies dq

Issues

1. EM has trouble converging for small disease allele frequencies dq

- This is due to uncertainty in resolving two component densities from a single component with a relatively large variance.

Issues

1. EM has trouble converging for small disease allele frequencies dq

- This is due to uncertainty in resolving two component densities from a single component with a relatively large variance.

2. Variances for each component don't converge to the expected common variance as expected in these extreme cases.

Issues

1. EM has trouble converging for small disease allele frequencies dq

- This is due to uncertainty in resolving two component densities from a single component with a relatively large variance.

2. Variances for each component don't converge to the expected common variance as expected in these extreme cases.

- I arbitrarily use a sample mean statistic across all three estimated mixture densities to estimate the common variance. This is only asymptotically unbiased (extremely small tolerance needed for rare minor alleles.)