PMR: Gaussians, Factor Analysis, Mixutres

Probabilistic Modelling and Reasoning

Amos Storkey

School of Informatics, University of Edinburgh

Outline

1 Gaussian

2 Factor Analysis

3 Gaussian Mixutre Models

Multivariate Gaussian

- $P(\mathbf{x} \in \mathcal{R}) = \int_{\mathcal{R}} p(\mathbf{x}) d\mathbf{x}$
- Multivariate Gaussian

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

- Σ is the covariance matrix

$$\Sigma = E[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T]$$

$$\Sigma_{ij} = E[(x_i - \mu_i)(x_j - \mu_j)]$$

- $\boldsymbol{\Sigma}$ is symmetric
- Shorthand $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- For $p(\mathbf{x})$ to be a density, $\boldsymbol{\Sigma}$ must be positive definite
- $\boldsymbol{\Sigma}$ has $d(d+1)/2$ parameters, the mean has a further d

Mahalanobis Distance

$$d^2_{\boldsymbol{\Sigma}}(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \mathbf{x}_j)$$

- $d^2_{\boldsymbol{\Sigma}}(\mathbf{x}_i, \mathbf{x}_j)$ is called the Mahalanobis distance between $\mathbf{x}_i$ and $\mathbf{x}_j$
- If $\boldsymbol{\Sigma}$ is diagonal, the contours of $d^2_{\boldsymbol{\Sigma}}$ are axis-aligned ellipsoids
- If $\boldsymbol{\Sigma}$ is not diagonal, the contours of $d^2_{\boldsymbol{\Sigma}}$ are *rotated* ellipsoids

$$\boldsymbol{\Sigma} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$$

  where $\boldsymbol{\Lambda}$ is diagonal and $\mathbf{U}$ is a rotation matrix
- $\boldsymbol{\Sigma}$ is positive definite $\Rightarrow$ entries in $\boldsymbol{\Lambda}$ are positive

Parameterization of the covariance matrix

- Fully general $\mathbf{\Sigma} \implies$ variables are correlated
- Spherical or isotropic. $\mathbf{\Sigma} = \sigma^2 I$. Variables are independent
- Diagonal $[\Sigma]_{ij} = \delta_{ij}\sigma_i^2$ Variables are independent
- Rank-constrained: $\mathbf{\Sigma} = \mathbf{W}\mathbf{W}^T + \mathbf{\Psi}$, with $\mathbf{W}$ being a $d \times q$ matrix with $q < d - 1$ and $\mathbf{\Psi}$ diagonal. This is the factor analysis model. If $\mathbf{\Psi} = \sigma^2 I$, then with have the probabilistic principal components analysis (PPCA) model

Transformations of Gaussian variables

- Linear transformations of Gaussian RVs are Gaussian

$$\mathbf{x} \sim N(\boldsymbol{\mu}_x, \boldsymbol{\Sigma})$$
$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{x}_0$$
$$\mathbf{y} \sim N(\mathbf{A}\boldsymbol{\mu}_x + \mathbf{x}_0, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)$$

- Sums of Gaussian RVs are Gaussian

$$Y = X_1 + X_2$$
$$E[Y] = E[X_1] + E[X_2]$$
$$\text{var}[Y] = \text{var}[X_1] + \text{var}[X_2] + 2\text{covar}[X_1, X_2]$$

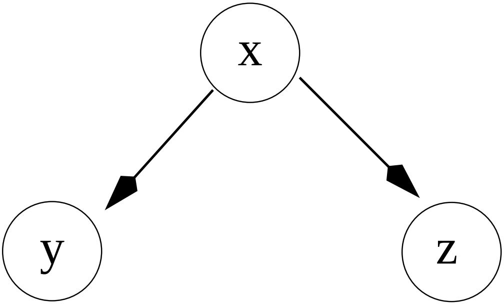
if X_1 and X_2 are independent $\text{var}[Y] = \text{var}[X_1] + \text{var}[X_2]$

Properties of the Gaussian distribution

- Gaussian has relatively simple analytical properties
- Central limit theorem. Sum (or mean) of M independent random variables is distributed normally as $M \to \infty$ (subject to a few general conditions)
- Diagonalization of covariance matrix $\implies$ rotated variables are independent
- All marginal and conditional densities of a Gaussian are Gaussian
- The Gaussian is the distribution that maximizes the entropy $H = -\int p(\mathbf{x}) \log p(\mathbf{x}) d\mathbf{x}$ for fixed mean and covariance

Graphical Gaussian Models

Example:



- Let X denote pulse rate
- Let Y denote measurement taken by machine 1, and Z denote measurement taken by machine 2.

- Model

$$X \sim N(\mu_x, v_x)$$
$$Y = \mu_y + w_y(X - \mu_x) + N_y$$
$$Z = \mu_z + w_z(X - \mu_x) + N_z$$

noise $N_y \sim N(0, v_y^N)$, $N_z \sim N(0, v_z^N)$, independent

- (X, Y, Z) is jointly Gaussian; can do inference for X given $Y = y$ and $Z = z$

As before

$$P(x, y, z) = P(x)P(y|x)P(z|x)$$

Show that

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_x \\ \mu_y \\ \mu_z \end{pmatrix}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} v_x & w_y v_x & w_z v_x \\ w_y v_x & w_y^2 v_x + v_y^N & w_y w_z v_x \\ w_z v_x & w_y w_z v_x & w_z^2 v_x + v_z^N \end{pmatrix}$$

Inference in Gaussian models

- Partition variables into two groups, $\mathbf{x}_1$ and $\mathbf{x}_2$

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}$$

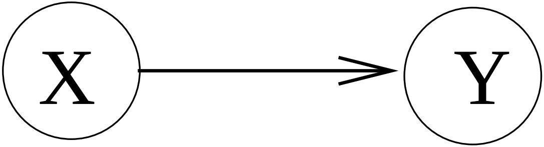
$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$$

$$\boldsymbol{\mu}^c_{1|2} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)$$

$$\boldsymbol{\Sigma}^c_{1|2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$$

- For proof see e.g. 2.3.1 of Bishop (2006) (not examinable)
- Formation of joint Gaussian is analogous to formation of joint probability table for discrete RVs. Propagation schemes are also possible for Gaussian RVs.

Example Inference Problem



$$Y = 2X + 8 + N_y$$

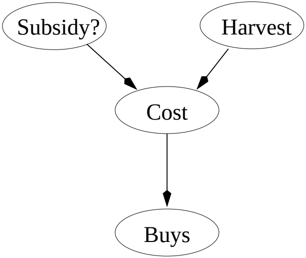
- Assume $X \sim N(0, 1/\alpha)$, so $w_y = 2$, $\mu_y = 8$, and $N_y \sim N(0, 1)$
- Show that

$$\mu_{x|y} = \frac{2}{4+\alpha}(y-8)$$
$$\text{var}(x|y) = \frac{1}{4+\alpha}$$

Hybrid (discrete + continuous) networks

- Could discretize continuous variables, but this is ugly, and gives large CPTs
- Better to use parametric families, e.g. Gaussian
- Works easily when continuous nodes are children of discrete nodes; we then obtain a *conditional Gaussian* model

Example



Model: Given that $Subsidy? = true$, cost c is a linear function of h, with a multiplication factor w_t and offset b_t, plus noise with variance v_t

$$P(Cost = c|Harvest = h, Subsidy? = true) \sim N(w_t h + b_t, v_t)$$

Similarly for $Subsidy? = false$

$$P(Cost = c|Harvest = h, Subsidy? = false) \sim N(w_f h + b_f, v_f)$$

Factor Analysis

- A latent variable model; can the observations be explained in terms of a small number of unobserved latent variables ?
- visible variables : $\mathbf{x} = (x_1, \ldots, x_d)$,
- latent variables: $\mathbf{z} = (z_1, \ldots, z_m)$, $\mathbf{z} \sim N(0, I_m)$
- noise variables: $\mathbf{e} = (e_1, \ldots, e_d)$, $\mathbf{e} \sim N(0, \Psi)$, where $\Psi = \mathrm{diag}(\psi_1, \ldots, \psi_d)$.

Assume

$$\mathbf{x} = \boldsymbol{\mu} + \mathbf{W}\mathbf{z} + \mathbf{e}$$

then covariance structure of $\mathbf{x}$ is

$$\mathbf{C} = \mathbf{W}\mathbf{W}^T + \boldsymbol{\Psi}$$

W is called the factor loadings matrix

$p(\mathbf{x})$ is like a multivariate Gaussian pancake

$$p(\mathbf{x}|\mathbf{z}) \sim N(\mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \boldsymbol{\Psi})$$

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}$$

$$p(\mathbf{x}) \sim N(\boldsymbol{\mu}, \mathbf{W}\mathbf{W}^T + \Psi)$$

- Rotation of solution: if $\mathbf{W}$ is a solution, so is $\mathbf{WR}$ where $\mathbf{RR}^T = I_m$ as $(\mathbf{WR})(\mathbf{WR})^T = \mathbf{WW}^T$. Causes a problem if we want to interpret factors. Unique solution can be imposed by various conditions, e.g. that $\mathbf{W}^T\mathbf{\Psi}^{-1}\mathbf{W}$ is diagonal.
- Is the FA model a simplification of the covariance structure? A full covariance has $d(d+1)/2$ independent entries. $\mathbf{\Psi}$ and $\mathbf{W}$ together have $d + dm$ free parameters (and uniqueness condition above can reduce this). FA model makes sense if number of free parameters is less than $d(d+1)/2$.

FA example

[from Mardia, Kent & Bibby, table 9.4.1]

- Correlation matrix

$$
\begin{array}{l}
\text{mechanics}\\
\text{vectors}\\
\text{algebra}\\
\text{analysis}\\
\text{statstics}
\end{array}
\qquad
\begin{pmatrix}
1 & 0.553 & 0.547 & 0.410 & 0.389\\
 & 1 & 0.610 & 0.485 & 0.437\\
 & & 1 & 0.711 & 0.665\\
 & & & 1 & 0.607\\
 & & & & 1
\end{pmatrix}
$$

- Maximum likelihood FA (impose that $\mathbf{W}^T\mathbf{\Psi}^{-1}\mathbf{W}$ is diagonal). Require $m \leq 2$ otherwise more free parameters than entries in full covariance.

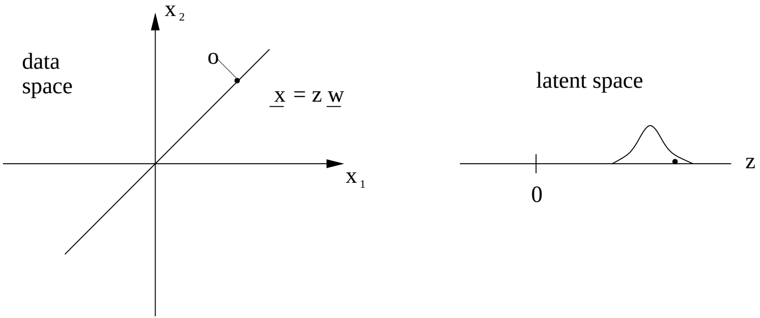
Variable	m = 1 $\mathbf{w}_1$	m = 2 $\mathbf{w}_1$	(not rotated) $\mathbf{w}_2$	m = 2 $\mathbf{w}'_1$	(rotated) $\mathbf{w}'_2$
1	0.600	0.628	0.372	0.270	0.678
2	0.667	0.696	0.313	0.360	0.673
3	0.917	0.899	-0.050	0.743	0.510
4	0.772	0.779	-0.201	0.740	0.317
5	0.724	0.728	-0.200	0.698	0.286

- 1-factor and first factor of the 2-factor solutions differ (cf PCA)
- problem of interpretation due to rotation of factors

FA for visualization

$$p(\mathbf{z}|\mathbf{x}) \propto p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$$

Posterior is a Gaussian. If $\mathbf{z}$ is low dimensional. Can be used for visualization (as with PCA)



Learning $\mathbf{W}$, Ψ

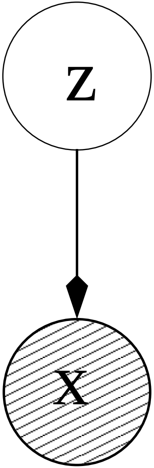
- Maximum likelihood solution available (Lawley/Jreskog).
- EM algorithm for ML solution (Rubin and Thayer, 1982)
    - E-step: for each $\mathbf{x}_i$, infer $p(\mathbf{z}|\mathbf{x}_i)$
    - M-step: do linear regression from $\mathbf{z}$ to $\mathbf{x}$ to get $\mathbf{W}$
- Choice of m difficult (see Bayesian methods later).

Comparing FA and PCA

- Both are linear methods and model second-order structure S
- FA is invariant to changes in scaling on the axes, but not rotation invariant (cf PCA).
- FA models *covariance*, PCA models *variance*

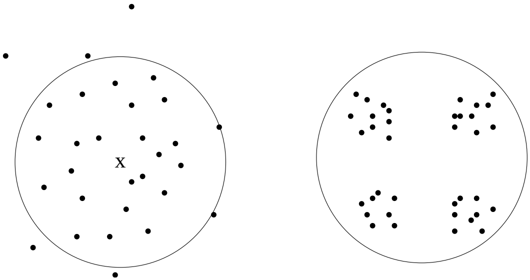
Hidden Variable Models

- Simplest form is 2 layer structure
- $\mathbf{z}$ hidden (latent) , $\mathbf{x}$ visible (manifest)
- Example 1: $\mathbf{z}$ is discrete $\rightarrow$ mixture model
- Example 2: $\mathbf{z}$ is continuous $\rightarrow$ factor analysis



Mixture Models

- A single Gaussian might be a poor fit



- Need mixture models for a *multimodal* density

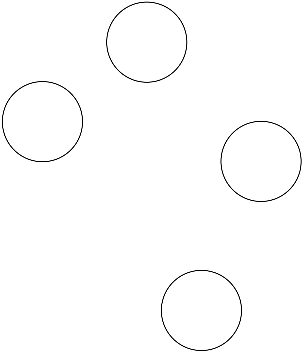
- Let $\mathbf{z}$ be a 1-of-k indicator variable, with $\sum_j z_j = 1$.
- $p(z_j = 1) = \pi_j$ is the probability of that the jth component is active
- $0 \leq \pi_j \leq 1$ for all j, and $\sum_{j=1}^{k} \pi_j = 1$
- The π_j's are called the *mixing proportions*

$$p(\mathbf{x}) = \sum_{j=1}^{k} p(z_j = 1) p(\mathbf{x}|z_j = 1) = \sum_{j=1}^{k} \pi_j p(\mathbf{x}|\theta_j)$$

- The $p(\mathbf{x}|\theta_j)$'s are called the mixture components

Generating data from a mixture distribution

for each datapoint
    Choose a component with probability π_j
    Generate a sample from the chosen component density
end for



Responsibilities

$$\gamma(z_j) \equiv p(z_j = 1|\mathbf{x}) = \frac{p(z_j = 1)\, p(\mathbf{x}|z_j = 1)}{\sum_\ell p(z_\ell = 1)\, p(\mathbf{x}|z_\ell = 1)}$$
$$= \frac{\pi_j\, p(\mathbf{x}|z_j = 1)}{\sum_\ell \pi_\ell\, p(\mathbf{x}|z_\ell = 1)}$$

- $\gamma(z_j)$ is the posterior probability (or responsibility) for component j to have generated datapoint $\mathbf{x}$

Max likelihood for mixture models

$$L(\theta) = \sum_{i=1}^{n} \ln \left\{ \sum_{j=1}^{k} \pi_j p(\mathbf{x}_i|\theta_j) \right\}$$

$$\frac{\partial L}{\partial \theta_j} = \sum_i \frac{\pi_j}{\sum_\ell \pi_\ell p(\mathbf{x}_i|\theta_\ell)} \frac{\partial p(\mathbf{x}_i|\theta_j)}{\partial \theta_j}$$

now use

$$\frac{\partial p(\mathbf{x}_i|\theta_j)}{\partial \theta_j} = p(\mathbf{x}_i|\theta_j) \frac{\partial \ln p(\mathbf{x}_i|\theta_j)}{\partial \theta_j}$$

and therefore

$$\frac{\partial L}{\partial \theta_j} = \sum_i \gamma(z_{ij}) \frac{\partial \ln p(\mathbf{x}_i|\theta_j)}{\partial \theta_j}$$

Example: 1-d Gaussian mixture

$$p(x|\theta_j) = \frac{1}{(2\pi\sigma_j^2)^{1/2}} \exp - \left\{ \frac{(x - \mu_j)^2}{2\sigma_j^2} \right\}$$

$$\frac{\partial L}{\partial \mu_j} = \sum_i \gamma(z_{ij}) \frac{(x_i - \mu_j)}{\sigma_j^2}$$

$$\frac{\partial L}{\partial \sigma_j^2} = \frac{1}{2} \sum_i \gamma(z_{ij}) \left[\frac{(x_i - \mu_j)^2}{\sigma_j^4} - \frac{1}{\sigma_j^2} \right]$$

At a maximum, set derivatives $= 0$

$$\hat{\mu}_j = \frac{\sum_{i=1}^n \gamma(z_{ij}) x_i}{\sum_{i=1}^n \gamma(z_{ij})}$$

$$\hat{\sigma}_j^2 = \frac{\sum_{i=1}^n \gamma(z_{ij})(x_i - \hat{\mu}_j)^2}{\sum_{i=1}^n \gamma(z_{ij})}$$

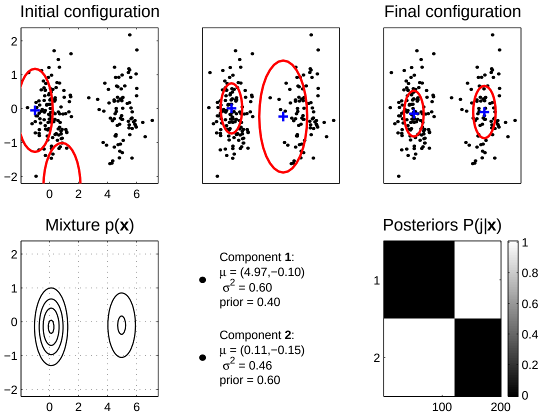
$$\hat{\pi}_j = \frac{1}{n} \sum_i \gamma(z_{ij}).$$

Generalize to multivariate case

$$
\begin{aligned}
\hat{\boldsymbol{\mu}}_j &= \frac{\sum_{i=1}^n \gamma(z_{ij})\mathbf{x}_i}{\sum_{i=1}^n \gamma(z_{ij})} \\
\hat{\Sigma}_j &= \frac{\sum_{i=1}^n \gamma(z_{ij})(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_j)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_j)^T}{\sum_{i=1}^n \gamma(z_{ij})} \\
\hat{\pi}_j &= \frac{1}{n}\sum_i \gamma(z_{ij}).
\end{aligned}
$$

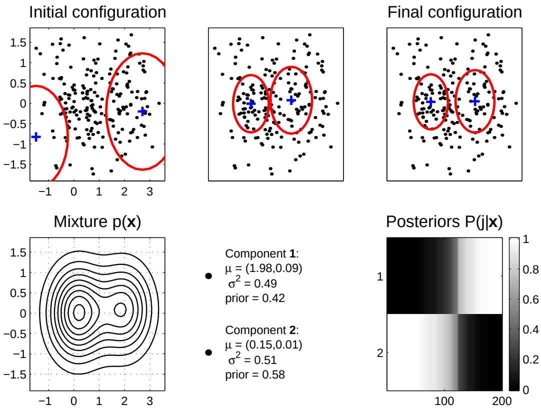
- What happens if a component becomes responsible for a single data point?

Example



(Tipping, 1999)

Example 2



(Tipping, 1999)

Kullback-Leibler divergence

- Measuring the “distance” between two probability densities $P(x)$ and $Q(x)$.

$$KL(P||Q) = \sum_i P(x_i) \log \frac{P(x_i)}{Q(x_i)}$$

- Also called the relative entropy
- Using $\log z \leq z - 1$, can show that $KL(P||Q) \geq 0$ with equality when $P = Q$.
- Note that $KL(P||Q) \neq KL(Q||P)$

The EM algorithm

- Q: How do we estimate parameters of a Gaussian mixture distribution?
- A: Use the re-estimation equations

$$\hat{\mu}_j \leftarrow \frac{\sum_{i=1}^{n} \gamma(z_{ij}) x_i}{\sum_{i=1}^{n} \gamma(z_{ij})}$$

$$\hat{\sigma}_j^2 \leftarrow \frac{\sum_{i=1}^{n} \gamma(z_{ij})(x_i - \hat{\mu}_j)^2}{\sum_{i=1}^{n} \gamma(z_{ij})}$$

$$\hat{\pi}_j \leftarrow \frac{1}{n} \sum_i \gamma(z_{ij}).$$

- This is intuitively reasonable, but the EM algorithm shows that these updates will converge to a local maximum of the likelihood

The EM algorithm

EM = Expectation-Maximization

- Applies where there is incomplete (or *missing*) data
- If this data were known a maximum likelihood solution would be relatively easy
- In a mixture model, the missing knowledge is which component generated a given data point
- Although EM can have slow convergence to the local maximum, it is usually relatively simple and easy to implement. For Gaussian mixtures it is the method of choice.

The nitty-gritty

$$L(\theta) = \sum_{i=1}^{n} \ln p(\mathbf{x}_i|\theta)$$

Consider for just one $\mathbf{x}$ first

$$p(\mathbf{x}|\theta) = \frac{p(\mathbf{x}, \mathbf{z}|\theta)}{p(\mathbf{z}|\mathbf{x}, \theta)}$$

so

$$\log p(\mathbf{x}|\theta) = \log p(\mathbf{x}, \mathbf{z}|\theta) - \log p(\mathbf{z}|\mathbf{x}, \theta).$$

Now take expectations wrt $p(\mathbf{z}|\mathbf{x}, \theta^{old})$

$$\log p(\mathbf{x}|\theta) = \sum_{z} p(\mathbf{z}|\mathbf{x}, \theta^{old}) \log p(\mathbf{x}, \mathbf{z}|\theta) - \sum_{z_i} p(\mathbf{z}|\mathbf{x}, \theta^{old}) \log p(\mathbf{z}|\mathbf{x}, \theta)$$

The nitty-gritty

$$L(\theta) = \sum_{i=1}^{n} \ln p(\mathbf{x}_i|\theta)$$

Consider for just one $\mathbf{x}_i$ first

$$\log p(\mathbf{x}_i|\theta) = \log p(\mathbf{x}_i, \mathbf{z}_i|\theta) - \log p(\mathbf{z}_i|\mathbf{x}_i, \theta).$$

Now introduce $q(\mathbf{z}_i)$ and take expectations

$$\begin{aligned}
\log p(\mathbf{x}_i|\theta) &= \sum_{z_i} q(\mathbf{z}_i) \log p(\mathbf{x}_i, \mathbf{z}_i|\theta) - \sum_{z_i} q(\mathbf{z}_i) \log p(\mathbf{z}_i|\mathbf{x}_i, \theta) \\
&= \sum_{z_i} q(\mathbf{z}_i) \log \frac{p(\mathbf{x}_i, \mathbf{z}_i|\theta)}{q(\mathbf{z}_i)} - \sum_{z_i} q(\mathbf{z}_i) \log \frac{p(\mathbf{z}_i|\mathbf{x}_i, \theta)}{q(\mathbf{z}_i)} \\
&:= \mathcal{L}_i(q_i, \theta) + KL(q_i||p_i)
\end{aligned}$$

From the non-negativity of the KL divergence, note that

$$\mathcal{L}_i(q_i, \theta) \leq \log p(\mathbf{x}_i|\theta)$$

i.e. $\mathcal{L}_i(q_i, \theta)$ is a *lower bound* on the log likelihood

We now set $q(\mathbf{z}_i) = p(\mathbf{z}_i|\mathbf{x}_i, \theta^{old})$ [E step]

$$\mathcal{L}_i(q_i, \theta) = \sum_{z_i} p(\mathbf{z}_i|\mathbf{x}_i, \theta^{old}) \log p(\mathbf{x}_i, \mathbf{z}_i|\theta) - \sum_{z_i} p(\mathbf{z}_i|\mathbf{x}_i, \theta^{old}) \log p(\mathbf{z}_i|\mathbf{x}_i, \theta^{old})$$
$$\overset{=}{def}\ Q_i(\theta|\theta^{old}) + H(q_i)$$

Notice that $H(q_i)$ is independent of θ (as opposed to θ^{old})

Now sum over cases $i = 1, \ldots, n$

$$\mathcal{L}(q, \theta) = \sum_{i=1}^{n} \mathcal{L}_i(q_i, \theta) \leq \sum_{i=1}^{n} \log p(\mathbf{x}_i|\theta)$$

and

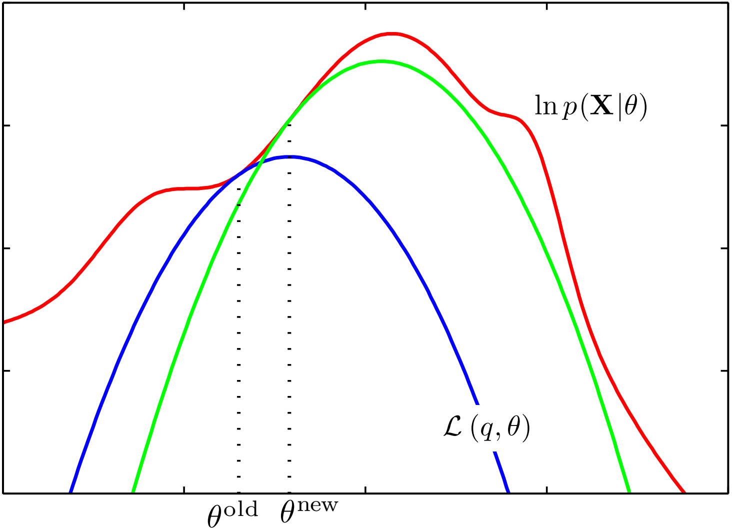
$$\mathcal{L}(q, \theta) = \sum_{i=1}^{n} Q_i(\theta|\theta^{old}) + \sum_{i=1}^{n} H(q_i)$$

$$\overset{=}{def}\; Q(\theta|\theta^{old}) + \sum_{i=1}^{n} H(q_i)$$

where Q is called the expected complete-data log likelihood. Thus to increase $\mathcal{L}(q, \theta)$ wrt θ we need only increase $Q(\theta|\theta^{old})$

Best to choose [M step]

$$\theta = \text{argmax}_\theta Q(\theta|\theta^{old})$$



$\ln p(\mathbf{X}|\theta)$
$\mathcal{L}(q,\theta)$
θ^{old}
θ^{new}

EM algorithm: Summary

E-step Calculate $Q(\theta|\theta^{old})$ using the responsibilities $p(\mathbf{z}_i|\mathbf{x}_i, \theta^{old})$

M-step Maximize $Q(\theta|\theta^{old})$ wrt θ

EM algorithm for mixtures of Gaussians

$$\mu_j^{new} \leftarrow \frac{\sum_{i=1}^n p(j|x_i, \theta^{old}) x_i}{\sum_{i=1}^n p(j|x_i, \theta^{old})}$$

$$(\sigma_j^2)^{new} \leftarrow \frac{\sum_{i=1}^n p(j|x_i, \theta^{old})(x_i - \mu_j^{new})^2}{\sum_{i=1}^n p(j|x_i, \theta^{old})}$$

$$\pi_j^{new} \leftarrow \frac{1}{n} \sum_{i=1}^n p(j|x_i, \theta^{old}).$$

[Do mixture of Gaussians demo here]

k-means clustering

initialize centres $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$
while (not terminated)
  for $i = 1, \ldots, n$
    calculate $|\mathbf{x}_i - \boldsymbol{\mu}_j|^2$ for all centres
    assign datapoint i to the closest centre
  end for
  recompute each $\boldsymbol{\mu}_j$ as the mean of the datapoints assigned to it
end while

k-means algorithm is equivalent to the EM algorithm for spherical covariances $\sigma_j^2 I$ in the limit $\sigma_j^2 \to 0$ for all j